PSEUDONYMISATION OF PERSONAL DATA ACCORDING TO THE GENERAL DATA PROTECTION REGULATION

Laura Tarhonen
Referee-artikkeli

1 Introduction

Pseudo-
Pretended and not real[1]

The General Data Protection Regulation (GDPR) sets out the rules regarding processing of personal data in the European Union.[2] This article will focus on the newly codified concept of pseudonymisation of data.[3] Pseudonymisation is processing of personal data in a manner that the data can no longer be linked to a specific data subject without the use of additional information. It seems to bring another layer to the definition of personal data but how should it be interpreted in practice? What qualifies as pseudonymisation? What is the relationship of pseudonymised personal data and anonymous information? When is

[1] See http://dictionary.cambridge.org/dictionary/english/pseudo (accessed 14 January 2017).

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

[3] The GDPR makes altogether 15 references to pseudonymisation. This is the first time it is defined in EU level data protection legislation since it was not defined in the Data Protection Directive. (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).

pseudonymisation useful and is it required by or incentivized in the regulation? This article will try to find answers to these questions.

2 The definition of pseudonymisation in the GDPR

Article four of the GDPR includes the definition of pseudonymisation. In this chapter I will break down the definition into elements and consider the relationships between pseudonymisation, anonymization and levels of de-identification. According to Article 4 of the GDPR:

> *'pseudonymisation' means*
> *a) the processing of personal data*
> *b) in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information,*
> *c) 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*[4]

2.1 The processing of personal data

The GDPR does not define pseudonymous data. According to the GDPR, there is only data that is personal data and data that is not. Also data that has undergone pseudonymisation is personal data and pseudonymising is just a way of processing data.[5] Perhaps the reason why the GDPR introduces a definition of "pseudonymisation" as opposed to "pseudonymised data" or "pseudonymous data" is, that the legislator wants to highlight that pseudonymisation should be considered as an activity and not as a static type of data.[6] For the sake of clarity, every time this article refers to "pseudonymised data" it means personal

---

[4] Breaks added by author.

[5] The concept of "processing" includes all activities and operations that can be done to personal data (see Article 4 (2).

[6] See UK's Information Commissioner Office's (ICO) analysis on the GDPR during the negotiations. https://ico.org.uk/media/1432420/ico-analysis-of-the-council-of-the-european-union-text.pdf (accessed 14 January 2017).

data that has undergone a pseudonymisation process as described in the GDPR. More specifically "pseudonymised data" is the part of the dataset that no longer can be attributed to specific data subject without additional information.

Recital 28 of the GDPR clarifies that pseudonymisation is not intended to preclude any other measures of data protection. Even if data is pseudonymised, the controller is not exempted from complying with data protection rules. Pseudonymisation does not seem to be incentivized by lifting any of the requirements of the regulation.

2.2 The personal data can no longer be attributed to a specific data subject without the use of additional data

Pseudonymisation is a processing activity that makes data no longer attributable to a specific data subject without the use of additional information, when that additional information is kept separately from the pseudonymised data.[7] Basically this means that when pseudonymising data, unique attributes are replaced by attributes from which the data subject can no longer be identified.[8] The replacing attribute can be either independent from the original value or derived from it, for example using hashing or encryption.[9] A simple example would be to replace data subject's first and last name[10] with a pseudonym that could be a random series of numbers or hashing the names with e.g. SHA-256 function that turns the names into 64 digits string of numbers and letters.

---

[7] In the definition of pseudonymisation, the term "personal data" is used when describing the data that is processed (pseudonymised) but when describing the additional data that is kept separate the word "information" is used. Typically "data" is perceived to be something raw and unstructured whereas "information" something more organized that is derived from a set of data.

[8] WP 29 Opinion 05/2014 on Anonymisation Techniques, p. 20.

[9] WP 29 Opinion 05/2014 on Anonymisation Techniques, p. 20.

[10] "Name" as a data type is a somewhat vague expression. In this article I use the word "name" as a higher level category that includes all available names. Typically datasets include first name and last name of data subject. The combination of names is a lot more identifying than first name or last name alone.

After pseudonymisation data cannot be attributed to a specific data subject without the use of additional information. This implies two things, first of which is that if data can be attributed to a specific data subject without the use of additional information then the data is not pseudonymised. Secondly, if the data cannot be attributed to a specific data subject with the use of additional data, then that data is not pseudonymised but rather anonymous information.[11]

2.2.1 Attributed to a specific data subject

What exactly does 'attributed to a specific data subject' mean? This question is the key to understanding, how pseudonymisation should be interpreted, since it defines on which level of identifiability the pseudonymised dataset can include personal data. The question can be divided to two parts: 1) definition of data subject and 2) interpretation of "a specific". The definition of data subject in the GDPR is "an identified or identifiable natural person". The meaning of identifiable natural person is further clarified: "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".[12]

Based solely on the definition of data subject, it seems that when pseudonymising data the controller should take into account both direct and indirect identifiers. This is because a data subject can be either identified or just identifiable natural person. Article 29 Working Party (WP29) has stated in its opinion about the concept of personal data, that concerning directly identified or identifiable persons, the name of the person is the most common identifier, and that in practice, the notion of identified person implies most often a reference to the person's name.[13] There are also other data types that can be categorized as directly identifying the data subject such as personal identification

[11] GDPR recital (26).

[12] GDPR Article 4(1).

[13] WP 29 Opinion 4/2007 on the concept of personal data p. 13.

number.[14] It seems to be clear that direct identifiers should be replaced when pseudonymising data; otherwise the data subject could be identified without additional information.

The question remaining is how to treat indirect identifiers and indirectly identifiable data. Indirect identifiability can refer to unique identifiers that can single out a specific data subject without directly identifying him or her (like a passport number), but it can also refer to unique combination of attributes that single out a data subject (like age, gender and region).[15] In practise, if pseudonymised data should no longer be attributable to an identified or identifiable person, would that mean that not only would pseudonymisation require replacement of all direct identifiers but also indirect identifiers and/or all data that could be used, in unique combinations, to indirectly identify data subject? This would lead to a very wide interpretation, and make pseudonymisation difficult since the controller would need to go through all the possible groups of attributes that together could be used to make a link between the pseudonymised data and a certain data subject.

Pseudonymised data cannot be attributed to a specific data subject without the use of additional information. The use of word "specific" seems to limit the scope of data types that would need to be replaced when pseudonymising data by demanding uniqueness. The data can only be linked to one specific data subject. All identifiers that can be linked to multiple data subjects could be left out when deciding which data types to replace. For example phone number is typically quite personal and points to a single data subject whereas address might not since there can be multiple people living in the same address.[16]

---

[14] See e.g. OECD Privacy Framework, 2013, p. 52 and *Mike Hintze:* Viewing the GDPR Through a De-Identification Lens: A Tool for Clarification and Compliance, 2016 p. 3, later (*Hintze 2016*) (https://fpf.org/wp-content/uploads/2016/11/M-Hintze-GDPR-Through-the-De-Identification-Lens-31-Oct-2016-002.pdf and see http://www.oecd.org/internet/ieconomy/privacy-guidelines.htm (both accessed 14 January 2017). Also during the preparation of Data Protection Directive social security number was seen as indirect identifier (WP 29 Opinion 4/2007 on the concept of personal data p. 12–13).

[15] See WP 29 Opinion 4/2007 on the concept of personal data p. 13–14.

[16] Phone numbers can be considered to be at least to some extent direct identifiers since typically people want to keep the same phone number and the telephone subscriptions are personal. When it comes to phone number as an identifier there might be cultural differences (e.g. in some countries phones can be shared) that

My interpretation of "attributed to a specific data subject" is that it means that the additional data kept separate from the pseudonymised data identifies directly a natural person or it can be used to do so.[17] This would mean that the minimum level of pseudonymisation could consist of replacing the names and personal identification numbers of the data subjects with pseudonyms. On the other hand, case by case consideration is needed, as many things in data protection depend heavily on the circumstances at hand. If the identification of a data subject would require also other types of data, than name, to make the identification more certain, those should also be replaced. These data types can include all unique identifiers, addresses, phone numbers et cetera. The wider possible interpretation could be that the additional information kept separate should include also the unique combinations of data that could indirectly identify a natural person.[18]

---

need to be considered but it is clear that as mobile devices get more and more personal the persistent identifiers the devices have also become more linked to a specific natural person.

[17] See e.g. WP 29 Opinion 4/2007 on the concept of personal data, p. 18: that describes pseudonymisation as "disguising identities"; ICO Anonymisation: managing data protection risk code of practice p. 49; that defines pseudonymisation as the process of distinguishing individuals in a dataset by using a unique identifier which does not reveal their 'real world' identity; Irish DPA: "Pseudonymisation" of data means replacing any identifying characteristics of data with a pseudonym, or, in other words, a value which does not allow the data subject to be directly identified. (https://dataprotection.ie/viewdoc.asp?DocID=1594&ad=1 published 2016; accessed 14 January 2017). In the WP 29 Opinion 05/2014 on Anonymisation Techniques pseudonymisation has been described as a technique of replacing one attribute (typically a unique one) in a record by other. It does give both an example of pseudonymisation where name, address and date of birth are replaced by a hash and also an example where just names are replaced. This would suggest that WP has seen pseudonymisation as technique something that can be used on "different levels" depending on controller's needs.

[18] The definition of data subject is "an identified or identifiable natural person" (GDPR Article 4) so when data is attributed to a data subject that data subject could be identified or identifiable from that data. In the German Federal Data Protection Act section 3 (6a) "aliasing" is defined as: replacing a person's name and other identifying characteristics with a label, in order to preclude identification of the data subject or to render such identification substantially difficult (http://www.gesetze-im-internet.de/englisch_bdsg/englisch_bdsg.html#p0008 accessed 14 January 2017). In the German context it is clear that also other data types need to be replaced than name but the definition seems to also be quite different from the GDPR's.

As a practical note, controllers have to remember, that person's name can often be visible in or a part of other data types, mainly email addresses, user names or nicknames. When considering how to pseudonymise a certain dataset, the controller should go through all data types and see if it is possible that data subject's name is stored in multiple data fields. Even if only a small number of data subjects have e.g. first name last name email address, it might be the easiest for the controller to replace all email addresses; user names (if not the same) and nicknames to be on the safe side.

The typical example of pseudonymisation has been changing names with pseudonyms or as Waltraut Kotschy put it: conversion of data about an identified person into data about a merely "identifiable" person.[19] I do think that there is a certain discrepancy between the wording of the definition of pseudonymisation in the GDPR and the way pseudonymisation as a technique has been described before the GDPR. This might be because the GDPR recitals do not clarify what exactly is expected from the controller when pseudonymising data and because it is not explicitly stated that pseudonymisation should not be seen as a fixed process but rather something that is subject to case by case consideration by the controller. Even though I believe that in some cases replacing data subject's name[20] from a dataset might already qualify as pseudonymisation, controllers are better off if they assess also the possibility to replace other unique attributes from the data.

2.2.2 Anonymous information

If pseudonymised data cannot be attributed to a specific data subject with the use of additional information then data is not pseudonymised but is anonymous. Anonymous data is not defined in the GDPR articles. The material scope of GDPR simply states that the regulation applies to processing of personal data. If data is not personal data then it is not in the scope of the regulation. This of course means that the definition

[19] *Waltraut Kotschy*, The new General Data Protection Regulation – Is there sufficient pay-off for taking the trouble to anonymize or pseudonymize data? 2016 p. 1 (https://fpf.org/wp-content/uploads/2016/11/Kotschy-paper-on-pseudonymisation.pdf accessed 14 January 2017).

[20] Including when applicable also personal identification numbers or social security numbers as well as other established unique identifiers.

of personal data plays a key role in defining applicable requirements. In short personal data means information that relates to an identified or identifiable natural person who can be identified either directly or indirectly.[21] GDPR recital 29 defines anonymous information as data that was never personal data (does not relate to an identified or identifiable natural person) or personal data that was rendered anonymous i.e. is no longer identifiable.[22]

WP 29 has highlighted in the opinion on anonymization techniques that pseudonymisation is not a method of anonymization and that it merely reduces the linkability of dataset with the original identity of a data subject.[23] Pseudonymised data shouldn't be confused to be anonymous data since the data subject can still be singled out of a group and hence identification is probable.

Recital 26 of GDPR describes pseudonymisation, identifiability of data subjects and data rendered anonymous. If pseudonymised data is such that it still can be attributed to a specific data subject by the use of additional information the data is information on identifiable natural person. In order to determine whether data subject is identifiable, controllers need to take into account all means that could reasonably likely, such as singling out, be used to identify the natural person directly or indirectly. Reasonable likelihood can be assessed through a) costs; b) amount of time; c) available technology at the time; and d) possible technical developments in the future.[24]

Recital 26 suggests that data could be rendered anonymous as a result of pseudonymisation. Presumably this could be the case if the controller would delete the additional data that could be used to re-identify a pseudonymised data. At least before the GDPR, Article 29 working party has however taken a strict view on anonymization. In its opinion on anonymization techniques WP 29 assessed anonymization techniques against three risks essential to anonymization: 1) singling out; 2) linkability; and 3) inference.[25] All three risks should be taken

---

[21] See GDPR Article 4.

[22] See also similar definition: WP 29 Opinion 4/2007 on the concept of personal data p. 21.

[23] WP 29 Opinion 05/2014 on Anonymisation Techniques, p. 3.

[24] GDPR recital (26).

[25] WP 29 Opinion 05/2014 on Anonymisation Techniques, p. 11. See also the

into account when anonymizing data. Anonymization that would produce truly anonymous data can be very difficult to achieve especially if aggregation of individual data rows would render the data useless. If it is not possible to address the risk of singling out (e.g. with aggregation of data), even efficiently mitigating the risks of linkability and inference might leave the controller uncertain of the anonymity of the data, especially because technologies are constantly developing and more and more data is collected in general. Even though the recitals of the GDPR do give controllers some help on how to assess if data is still identifiable, it will still be difficult for the controllers to define when data has been anonymized properly and whether that can happen as a result of a pseudonymisation process.

2.3 Additional information is kept separately and is subject to technical and organisational measures

According to the definition of pseudonymisation, the additional data that could be used to attribute pseudonymised data to a specific data subject should be kept organisationally and technically separate from the pseudonymised dataset. The controller can still be able to re-identify the data and it is not explicitly required that the link between the pseudonymised dataset and the keys to re-identification[26] should be deleted. However, as long as the keys to re-identification are saved after pseudonymisation, the controller cannot claim that the data is anony-

---

guidance of Irish Data Protection Authority: https://dataprotection.ie/viewdoc.asp?DocID=1594&ad=1 (accessed 14 January 2017). I will not go into details in explaining these risks but shortly: singling out means the possibility to isolate some or all records which identify an individual in the dataset meaning that unique records can be identified from the data; linkability, means the ability to link, at least, two records concerning the same data subject or a group of data subjects meaning that data is organized around identifiers that are linked together; inference means the possibility to deduce, the value of an attribute from the values of a set of other attributes meaning e.g. that by combining multiple data types it might be possible to deduct data that identifies the data subject.

[26] By the "keys to re-identification" I mean the combination of the identifiers in the pseudonymised dataset and the additional information than can be used to reverse the pseudonymisation.

mous.[27] The technical and organisational measures used during pseudonymisation must ensure that the pseudonymised data is not combined with the keys to re-identification. Technical measures can be measures that directly involve IT systems and in this context technical measures could mean e.g. storing the pseudonymised data securely in a different database or storing environment, than the keys to re-identification. Technical measures could also include encryption of data. Organizational measures relate to physical environments as well as to the people using and having access to certain IT systems. The controller should have documented access rights to both the pseudonymised dataset and to the keys.[28] If only a limited number of assigned people, or no one, has the access to both of the datasets, the controller can demonstrate that it is complying with the separation requirement and has created a "Chinese wall" between the two datasets. Organizational measures can also refer to contractual measures e.g. when controller outsources data processing to a processor and contractually restricts the re-identification of data.[29]

It is somewhat unclear whether pseudonymisation is meant to be an irreversible measure. The same controller can both process the pseudonymised dataset and the additional information that could be used to attribute the data to specific data subjects but the definition of pseudonymisation does require that controller ensures that the pseudonymised data is not attributed to an identified or identifiable natural person. This would imply that pseudonymisation is intended to be irreversible even though it would still be possible to link the additional data to the pseudonymised dataset. I think that the reversibility needs to be considered case by case taking into account the processing purposes and why pseudonymisation is applied to the data in the first place.

---

[27] If however the keys to re-identification are deleted, it is possible that the data is rendered anonymous. So it is possible that the controller first pseudonymises a dataset still storing the keys to re-identification and as a separate measure later on anonymizes the data by deleting the keys.

[28] See GDPR recital (29).

[29] *Khaled El Emam, Eloise Gratton, Jules Polonetsky, and Luk Arbuckle*: The Seven States of Data: When is Pseudonymous Data Not Personal Information? 2016 page 9 (Later: *El Emam, Gratton, Polonetsky, Arbuckle 2016*) https://fpf.org/wp-content/uploads/2016/11/El-Emam_States-of-Data-Main-Article-short-v6.pdf (accessed 14 January 2017).

If pseudonymisation is used as a security measure for e.g. customer or similar databases, the pseudonymisation needs to be reversible in order it to be useful for most controllers.[30] If controller pseudonymises its customer database so that directly identifiable data is kept separate from other data, like purchase history, the controller may still need to combine the databases for many legitimate purposes such as delivery of service and marketing. In this type of situation pseudonymisation could be seen as a "mode of storage".

If pseudonymisation is used as a data minimization measure the keys to re-identification should be deleted. This could be the case when controller no longer needs to process data in an identified form but could use it for example for statistical or reporting purposes in a pseudonymised form. In this type of situation pseudonymisation is used as a data protection by design practice.

2.4 Different levels of de-identified data

Pseudonymisation is defined as a processing activity in which personal data that previously could have been attributed to a specific data subject no longer can be. Does this mean that pseudonymisation is only a process for datasets that permit the direct identification of the data subject in the first place? Or is it possible to pseudonymise also datasets that by their nature are not directly linkable to a specific data subject, without the use of additional data, if the dataset would be made even less identifiable? These questions could be important if pseudonymising would benefit the controller when assessing the compliance with the requirements of the GDPR.[31]

---

[30] One can also argue that pseudonymisation must be reversible for it to be more useful than anonymization. For example in medical research it might be necessary to contact or change patients treatment as a result of the research.

[31] This might be more a theoretical discussion than a practical one because the harder it is to link data to specific data subject the smaller the risks for individuals are and this can be taken into account when applying GDPR's requirements to the data. Therefore making data less identifiable data should always be considered as a data protection by design and security measure. But nevertheless it is important that controllers understand how the definitions of the regulation should be reflected in their thinking.

It does not seem like the definition of pseudonymisation in GDPR would have been intended to extend to data that is already when collected, difficult to link directly to a specific data subject without the use of additional data. The definition of pseudonymisation includes the idea that the personal data can no longer be attributed to a specific data subject. This would imply that personal data that was never attributable to a specific data subject, without the use of additional data, could not be pseudonymised. An example of data that is by nature difficult to link to a specific data subject is online behavioural data that is normally collected via cookies and linked to an IP-address. Both the cookie id and IP-address in most cases are personal data but by nature they can be connected to identifiable natural person indirectly.[32] Additional information would be needed to identify a specific data subject and often the controller does not have the data needed to do the identification. For example IP-addresses can normally only be linked to a natural person by their internet service providers.[33]

Pseudonymised data can be defined as a subset of identifiable data.[34] The identity of the data subject cannot be directly distinguished from the dataset but if additional information is used the data subject can be identified. The Article 11 of the GDPR suggests that the regulation recognizes the need for multiple levels of identifiability. If the purposes for which controller processes personal data do not require identification of data subject, the controller is not obliged to processes additional information just to identify the data subject for the purpose of complying with the regulation. If the controller is able to demonstrate that it is not in a position to identify the data subject and the data subject cannot provide additional information enabling identification, then the controller does not need to carry out data subject's rights as set out in the GDPR. It seems that the Article 11 takes de-identification of data one step further than pseudonymisation, since even additional data could not identify the data subject. This supports the arguments,

[32] The GDPR explicitly mentions online identifiers as personal data that can be directly or indirectly used to identify a natural person. See GDPR Article 4(1).

[33] This is not to say that controllers could not sometimes link IP-address to a specific identified data subject but this would require also other data than IP-address. IP-address as personal data see: Scarlet Extended (C-70/10, EU:C:2011:771) and Breyer (C582/14, EU:C:2016:779).

[34] *Hintze 2016*, p. 3.

made in the section 2.1.1, that pseudonymisation is intended as measure that disguises the "direct identity" of a data subject and does not necessarily go so far as reducing the indirect identifiability; otherwise Article 11 would probably make a reference to pseudonymisation.[35]

As some of the data subjects rights might be burdensome to comply with, Article 11 can encourage controllers to de-identify data when it is possible. This could also be seen as extension of the minimum pseudonymisation of replacing direct identifiers with pseudonyms. If the controller designs a pseudonymisation process that goes further than the minimum and replaces or deletes also easily linkable indirect identifiers (address, phone number, date of birth etc.), it might achieve a level of de-identification referred to in Article 11. Of course from practical point of view, applying Article 11 is only possible when the identification of data subject is not necessary whereas in reality often the personal data is needed in an identifiable form in order to simply provide the intended services to data subjects. There can however be datasets, like collecting online behavioural data, that could be usable for the controller but it is not necessary to maintain the link between the identifiable data subject and the data.

3 Pseudonymisation in other articles of the GDPR

A part from the Article 4 on definitions, pseudonymisation is mentioned in five other articles of the GDPR. In this chapter I will look into these articles and the way pseudonymisation is linked to actual requirements of the regulation.[36]

[35] In the WP29 Guidelines on the right to data portability (published 13 December 2016) WP 29 clarifies that also pseudonymous data that can be clearly linked to a data subject e.g. by data subject providing an identifier is well within the scope of data portability (See especially page 7). What is interesting in this guidance is that WP29 has chosen to use the term "pseudonymous data" not the term "pseudonymised data". It is therefore somewhat unclear whether they refer to data that was pseudonymised according to the GDPR or simply to pseudonymous data (that is not defined in GDPR).

[36] I will not consider the Article 40 about codes of conduct in as it merely states that pseudonymisation is one topic that could be addressed in sector specific codes of conduct which would clarify how the regulation should be properly applied.

No.	Article	Recital	Context
4	Definitions	26, 28, 29	Defines what is pseudonymisation
6	Lawfulness of processing		Pseudonymisation an example of a safeguard that can be taken into account when assessing compatibility of new data processing purposes in the purpose limitation test
25	Data protection by design and by default	75, 78	Pseudonymisation as a measure that demonstrates the implementation of data protection principles
32	Security of processing	85	Pseudonymisation as an information security measure
40	Codes of conduct		Pseudonymisation as a topic to cover in codes of conduct
89	Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes	156	Pseudonymisation as an example of safeguard which demonstrates that data minimisation principle is respected

Table 1. Pseudonymisation in the GDPR Articles.

3.1 Lawfulness of processing and purpose limitation (Article 6)

Article 6 of the GDPR lists the legal bases under which, processing of personal data is lawful. In short the legal bases are: consent, contract, legal obligation, vital interest, public interest task and legitimate interest. The Article 6 does not include any direct exemptions, reliefs or requirements for pseudonymisation of data. During the negotiations of the GDPR it was proposed that the definition of legitimate interest in Article 6, would include that pseudonymisation of data should be taken into account when assessing the legitimate interest of controller since the controller has minimized the risks the processing poses to data sub-

jects.[37] The proposal did not make it to the final regulation text but the recital 28 confirms in general that pseudonymisation can reduce the risks to the data subjects concerned and help controllers and processors to meet their data protection obligations. In practise controllers can use pseudonymisation of data as an argument to support their legitimate interest for data processing, but merely pseudonymising data is not necessarily enough to demonstrate that the legitimate interest is not overridden by the interests or rights of the data subject.

Personal data can only be collected for specified, explicit and legitimate purposes and it should not be further processed in a manner that is incompatible with the original purposes.[38] If a controller wants to process data to a purpose other than for which it was originally collected, and the controller does not have data subjects consent (or other legal grounds) for the processing, the controller needs to do a "purpose limitation balancing test"[39] to make sure that the new purpose is compatible with the original purposes. When assessing if the new purpose is compatible with the original purposes, pseudonymisation as a safeguard can be considered as mitigating the risks for data subject's data protection.[40] As an example: an e-commerce company has a customer database that includes purchase history and behavioural data of how customers interact with their web service. The company has originally collected the data for delivery and development of the service. It has informed its customers about how data is processed in a privacy notice. Now the company is planning to launch a totally new web service and would like to use the data they have in the customer database to do product development of the new service. If they would be able to demonstrate that the data in the customer database is sufficiently pseudonymised, that could be used to argument that the further processing of data is compatible with the original purposes of processing. When assessing if pseudonymisation is an appropriate safeguard in the spe-

[37] This was suggested by the German delegation during the negotiations of GDPR. See: http://www.statewatch.org/news/2015/jan/eu-council-dp-reg-pseudonymisation-14705-rev1-14.pdf (accessed: 14 January 2017) See also: *El Emam, Gratton, Polonetsky, Arbuckle 2016* especially pages 6–8.

[38] GDPR Article 5.

[39] Outlined already in WP 29 Opinion 3/2013 on purpose limitation, p. 3 and now in the Article 6 (4).

[40] GDPR Article 6 (4)(e).

cific data processing situation, authorities could take into account for example the level of pseudonymisation and whether any anonymization techniques are also used to make the data less identifiable; is it possible to re-identify the pseudonymised data; and is there technical and / or organisational means in place to make the pseudonymisation more effective like encryption or restricted access rights.

3.2 Processing of personal data for public interest, scientific or historical research purposes or statistical purposes (Article 89)

Personal data should not be stored for longer than necessary for the purposes for which the data is processed. The principle of storage limitation is one of the six basic principles of EU data protection law. The GDPR makes an exemption on the storage limitation if the data is processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes and appropriate technical and organisational measures are taken to safeguard the rights and freedoms of the data subject.[41]

Article 89 specifies rules regarding processing of data for public interest archiving, scientific or historical research and statistical purposes. The controller must have safeguards in place that ensure respect for the principle of data minimization.[42] As data can be stored longer for these purposes, that serve the "greater good", it is important that in each situation the controller assesses critically which data is necessary for the purposes of processing and deletes all data that is not. Article 89 encourages the use of pseudonymisation as a safeguard, while still admitting that it might not be possible in all cases because pseudonymisation might not serve the original purpose. This might be the case e.g. when it is necessary to archive public records per identified data subject. However in most cases, pseudonymisation could probably be

[41] GDPR Article 5. Article 5 includes also an exemption of purpose limitation when data is further processed for public interest archiving, scientific or historical research or statistical purposes if data is processed in accordance with Article 89(1).

[42] GDPR Article 89.

used at least as a security measure when storing the data even if it would be necessary to reverse the pseudonymisation from time to time.

If it is possible for the controller to fulfil its processing purposes with data that no longer permits the identification of data subjects, the controller needs to process the data in that manner.[43] From the Article 89 text it is unclear whether this means that data should be anonymized (no longer personal data) or if pseudonymisation (data subject cannot be identified without additional information) is enough. The controller should assess case by case the level of identifiability necessary and the safeguards that should be used. The bottom line seems to be that the exemption of storage limitation in Article 5 together with Article 89 should not be read as blanket rules that data can be stored forever for public interest archiving, scientific or historical research or statistical purposes. The controller should choose to use the least intrusive means of processing data and pseudonymising can be used to demonstrate this.

3.3 Data protection by design and by default (Article 25)

Data protection by design and default has been codified in the GDPR as a requirement for controllers of personal data. Data protection by design and default is an approach to processing of data where the controller plans and executes the data processing privacy friendly from the start as a default.[44] It has sort of already been present in the data protection principles of the Data Protection Directive, or it can at least be read between the lines, but now it has been specified as a requirement.[45] The Article 25 does not explicitly define how the controller needs to build data protection into the data processing activities. The requirement is proportional meaning that each data processing needs to be assessed separately and the controller has to take into account the nature, scope, context and purposes of processing as well as the risks

[43] GDPR Article 89 (1).

[44] See for example ICOs explanation on privacy by design and its benefits: https://ico.org.uk/for-organisations/guide-to-data-protection/privacy-by-design/ (accessed: 14 January 2017).

[45] Data Protection Directive (95/46/EC), Article 6.

for data subject's rights and freedoms. Pseudonymisation is mentioned as technical and organisational measure that can be designed to implement data protection principles such as minimization. In general when pseudonymisation is linked to data minimization, it means that only necessary data is processed and unnecessary data is deleted e.g. the keys to re-identification should be deleted.

It is likely that controllers will need to demonstrate how they have implemented data protection by design into their processing activities. When assessing if the controller has fulfilled the requirement, the existence of pseudonymisation process for personal data can be used to argument that the controller has tried to minimize the risks the processing activities might have. Also since pseudonymisation has been explicitly mentioned in the Article 25, it is likely that authorities would ask whether or not controllers have used pseudonymisation when processing personal data. Still it is good to keep in mind that data protection by design and default is also more than just pseudonymising data. It is about designing the whole data processing life cycle in a manner that respects the data protection principles in the context of that specific data processing. As an example if the controller pseudonymises a dataset by just replacing direct identifiers like the data subject's names with pseudonyms, but the controller's processing purposes could be fulfilled also with less data e.g. without all indirect identifiers of the dataset, it could be argued that the pseudonymisation the controller did was not enough to comply with the data protection by design and default requirement.

3.4 Security of processing (Article 32)

In order to safeguard personal data the controller of data must implement appropriate technical and organisational measures to ensure the security of personal data. The measures should be in proportion with the risks that the nature, scope, context and purposes of processing pose to the rights and freedoms of natural persons. Similarly to the requirement of data protection by design and default, security of processing can be assessed against what is proportionate in each case taking into account the riskiness of the processing to natural persons. It is clear that if a controller processes for example health data or

credit card numbers more emphasis needs to be put to the security of the data to prevent e.g. accidental loss or unauthorized disclosure of the data.

One of the security measures mentioned in the Article 32 is pseudonymisation of personal data. Recital 78 demands that controllers should, in order to be able to demonstrate compliance with the regulation, adopt internal policies and implement measures such as pseudonymising data as soon as possible. "As soon as possible" could refer to pseudonymisation already when collecting data but it could also refer to minimizing data during processing activities. Having a process in place that would pseudonymise data at the time of collection, would from security point of view decrease the risk of identifying data subjects since there would be no backups or other temporary storages containing the whole dataset together. This of course applies only if the data processing purposes require re-identification of data because otherwise the controller should not be processing identified data at all.[46]

Strictly from data security point of view pseudonymisation can be regarded as a measure to limit the risks that for example a data leak could have but pseudonymisation can also be seen as a measure that enables the controller to define appropriate access rights for internal and external data processors. If for example an employee would not necessarily need access to the identified data, they could only be assigned with access to pseudonymised data. This applies to outsourcing situations as well. If processor would not need access to the whole dataset controller could limit its risks by just assigning the processor access to pseudonymised data. The aforementioned measures could also demonstrate data protection by design. It can actually be quite difficult to distinguish between security safeguards and data protection by design measures but the division is probably not even necessary since both measures contribute to the same thing – processing data in a lawful and appropriate manner. Like with data protection by design and default requirement, pseudonymisation should not be seen as the

[46] If the purposes for which the data is collected do not require re-identification of the data subject, the controller should not collect e.g. the name of the data subject in the first place at all and no separate "pseudonymisation" would be needed.

only safeguard that the controller needs to put in place to comply with data security requirement.

One of the key risks in data processing is a personal data breach.[47] There are many different types of data breaches, but a typical one would be such where personal data is accessed by or disclosed to an unauthorized party. Properly pseudonymised data can limit the risks that these types of breaches could have to natural persons. Especially in the cases where either the additional data that could be used to re-identifying data subjects is not breached or it has been deleted altogether.[48] However if data is pseudonymised merely by replacing the names of data subjects with pseudonyms, the risk that the third parties can link the data to identified individuals is probable and pseudonymisation wouldn't provide significant protection for data subjects.

In an ideal world the controller would always use the best security techniques available when processing personal data. This is however not always possible and therefore pseudonymisation may prove to be a good factor in prioritizing assets that need to be protected. For example if controller has limited funds to allocate to information security measures, it can focus on the databases that contain the most sensitive or identifiable personal data, like the keys to re-identify pseudonymised dataset, and use strong encryption to safeguard those databases.

4 Conclusions

In this article I have been answering the questions: what is pseudonymisation and does GDPR require it or incentivize its use. In a nutshell pseudonymisation is a process of de-identifying identified data where the keys to re-identification are kept separate, deleted or found elsewhere from the pseudonymised data. Personal data that has been pseudonymised is still personal data and data protection

[47] See recital (85) for examples of risks to the data subject.

[48] Strong encryption of the identified data ("keys") can further reduce the risk since it might not be possible for the unauthorised party to decrypt the dataset even if they would gain access both pseudonymised data and the keys of re-identification.

rules fully apply to it. If it is necessary to classify datasets that have been pseudonymised those could be described as datasets containing identifiable data.

Pseudonymisation is not a method of anonymization but it can facilitate anonymization if used with other techniques like generalization and deletion of data. Pseudonymisation, according to the GDPR, may not have been intended to have multiple levels of identifiability, but as a method pseudonymisation can definitely be used to achieve proportionate and sufficient levels of identifiability to cater to the needs of controller's processing purposes, while also respecting the rights and freedoms of natural persons.

As a concept pseudonymisation is not new. Still pseudonymisation is one of those concepts that need clarification from authorities and case law in order for the controllers to be able to assess its true usability and the incentives to use it. So far it is difficult to come up with direct benefits for using pseudonymisation except that it demonstrates lawfulness of processing. The data protection authorities and courts can incentivize the use of pseudonymisation in their guidance and judgements e.g. by imposing smaller administrative fines for controllers that have had established pseudonymisation practices. Many controllers could also benefit from "best practices" and sector specific codes of conduct addressing the practicalities of pseudonymisation. Also the emergence of specialized third party service providers that would perform pseudonymisation on behalf of controllers, and for example store the keys to re-identification securely, could be accelerated by clear instructions from self-regulation or authorities.

Especially in rapidly changing online context, it can be difficult for the controller to specify explicit purposes for data processing before the processing even begins. Well thought pseudonymisation process can help the controller demonstrate that it has grounds to further process data to also for other purposes than the ones which data was originally collected for. Of course this applies only if the new purposes are still compatible with the original ones. In these types of situations it is clear that pseudonymisation can be beneficial for the controller.

In the GDPR, pseudonymisation is referred to as a measure or an example of a way to safeguard personal data. Whether it is explicitly required by the GDPR or not, is subject to debate. I would argue that

it is required when it is a suitable measure in connection with specific data processing but that a controller can be compliant with the GDPR also without pseudonymising data.

In order to comply with data protection by design and default and data security requirements just pseudonymising data is not necessarily enough. Pseudonymisation seems to be most usable in data processing where it is not necessary constantly re-identify the data subjects and hence the keys for re-identification can be deleted. If controller typically has the need to process data in connection with identified data subject, pseudonymisation should be seen as data protection by design and data security measure but it should be acknowledged that especially as data security measure it can be quite limited.

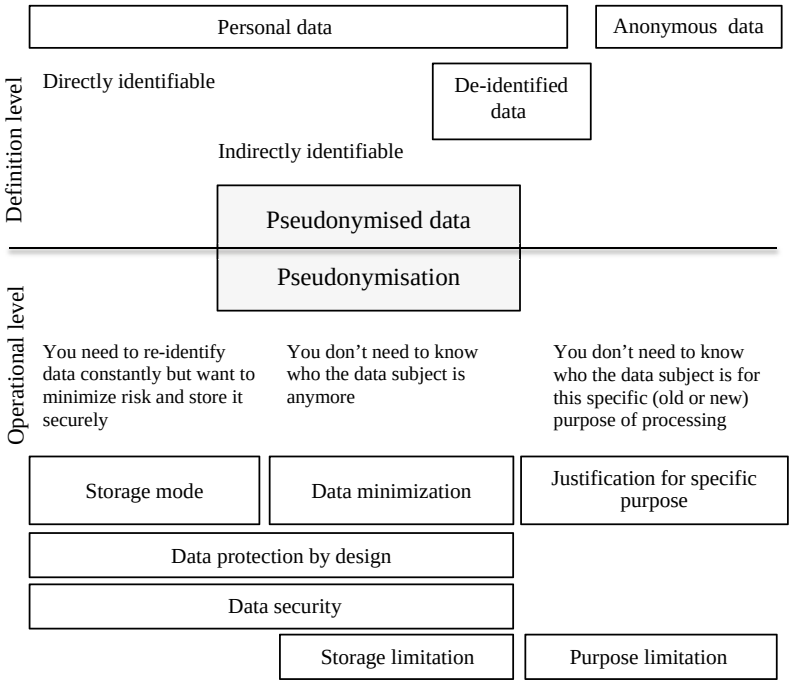

Figure 1. Overview of pseudonymisation.

I started this article with the definition of the prefix “pseudo”. It is used when something is not real and is pretended. Pseudonymisation of data does seem to mean just that. Changing some parts of the data, so that it is not true and is pretended: the name field of data subject is replaced by a number pretending to be the name. The value of pseudonymisation according to GDPR might also be a bit pretentious. As controllers prepare for the application of GDPR and start to refine their data protection processes, they should not expect pseudonymisation to be the answer to all of their problems. Pseudonymised data is still personal data and there are not too many shortcuts to compliance. My message to all controllers is that fitting pseudonymisation measures into your personal data processing is a part of the journey to lawfulness, but it should not be relied on too much. Controllers need to actively assess whether pseudonymisation will, in the context of their data processing, be a usable and sufficient safeguard either from data minimization (data protection by design and default) or security point of view. In many cases controllers will find that it is not and that they need to pursue also more extensive de-identification measures.