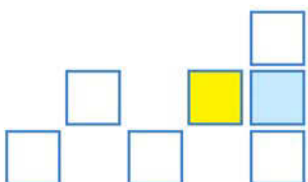


# Towards a Better Understanding of Identification, Pseudonymization, and Anonymization



Unabhängiges Landeszentrum für  
Datenschutz Schleswig-Holstein



Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein (ULD), 2021

With the exception of the public domain picture of figure 36, the project logos and the ULD logo, both text and figures of the present report are © 2021 by the Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and are licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>). Please include the source location <https://uld-sh.de/PseudoAnon> in your attribution.

## Acknowledgements

Constant feedback from and review by Marit Hansen, Harald Zwingelberg, Benjamin Bremert, all ULD, and by Jessica Schroers, KU Leuven, significantly helped in shaping and improving the content.

Parts of the underlying research and writing have been conducted as part of the projects *PANELFIT*<sup>1</sup> and *Forum Privatheit*<sup>2</sup>.



PANELFIT project has received funding under the European Union's H2020 research and innovation programme under grant agreement No 788039.



Forum Privatheit has received funding from the German Federal Ministry of Education and Research (BMBWF) – Förderkennzeichen for ULD: 16KIS0747.

---

<sup>1</sup> <https://panelfit.eu>

<sup>2</sup> <https://www.forum-privatheit.de/>

**Table of Contents**

- 1 Objective and Approach ..... 7**
- 2 Outline..... 7**
- 3 The Concept of Identification ..... 7**
  - 3.1 Information..... 8
  - 3.2 Information Elements with Relevance to Identification..... 9
    - 3.2.1 Uniqueness and the Dimension of the Data Set..... 15
  - 3.3 Individual-Level Information and Data Records ..... 16
  - 3.4 Linking of Information ..... 16
    - 3.4.1 Special kinds of Linking..... 20
  - 3.5 Identification ..... 22
  - 3.6 Risk of identification..... 26
  - 3.7 Measures to Reduce the Risk of Identification..... 29
    - 3.7.1 Restricting Access to Personal Data ..... 29
    - 3.7.2 Avoiding access by any actors motivated to identify data subjects .. 29
    - 3.7.3 Avoiding access by any actors who may already have information about data subjects ..... 29
    - 3.7.4 Prevent the coming together of the personal data with additional information..... 30
    - 3.7.5 Reducing the identification potential of the data itself..... 30
      - 3.7.5.1 Deterministic linking of unique handles..... 30
      - 3.7.5.2 Linking of quasi-identifiers ..... 32
      - 3.7.5.3 Linking of identity-relevant properties ..... 34
      - 3.7.5.4 Summary of transformations that reduce the identification potential of data..... 43
      - 3.7.5.5 Tools for reducing the identification potential of personal data 45
- 4 Pseudonymization..... 46**
  - 4.1 Introduction to pseudonymization ..... 46
    - 4.1.1 Motivation to use pseudonymization ..... 46
    - 4.1.2 Risk reduction through pseudonymization ..... 48
    - 4.1.3 Importance of Pseudonymization in the GDPR ..... 49
  - 4.2 Pseudonymization in the GDPR ..... 50
  - 4.3 The context of pseudonymization and access to additional information..... 52
  - 4.4 Usage scenarios of pseudonymization ..... 53

4.5	Definition of concepts relevant to pseudonymization.....	55
4.5.1	Data pseudonymization.....	55
4.5.2	Re-identification.....	56
4.5.2.1	(General) re-identification.....	56
4.5.2.2	Planned re-identification.....	57
4.5.3	(Directly) identified (personal) data.....	57
4.5.4	Pseudonymous data .....	58
4.5.4.1	(General) pseudonymous data.....	58
4.5.4.2	Strictly pseudonymous data.....	58
4.5.4.3	Pseudonymized data .....	61
4.5.5	Additional information .....	61
4.5.5.1	(General) additional information .....	61
4.5.5.2	Split-off additional information.....	62
4.5.5.3	Different types of additional information.....	62
4.5.6	Pseudonyms .....	65
4.5.6.1	(General) pseudonyms .....	65
4.5.6.2	Pseudonymous handles .....	65
4.6	Data pseudonymization in detail .....	66
4.7	Technical and organizational measures for pseudonymization .....	69
4.8	Different types of (re-) identification.....	73
4.9	Pseudonymization and Art. 11 GDPR.....	75
4.9.1	Different types of additional information during pseudonymization.....	76
4.9.2	Identifying data subjects with different types of split-off additional information.....	78
4.9.3	Additional information provided by the data subject.....	80
4.9.4	Trustworthiness of additional information provided by data subjects	82
4.9.5	Pseudonymous Credentials .....	83
4.9.6	Summary of identifiability of data subjects and informing data subjects	85
4.9.7	Waived obligations due to inability to identify.....	86
4.9.8	Data subject rights and the need for re-identification .....	87
<b>5</b>	<b>Anonymization.....</b>	<b>91</b>
5.1	Definition of Anonymous .....	91
5.2	Comparison of anonymous with strictly pseudonymous data .....	95

5.3	Concepts relevant to anonymization .....	97
5.4	Functional description of (successful or attempted) anonymization .....	98
5.5	Do anonymous data exist? .....	99
5.6	Concepts relevant to the identifiability of data .....	100
5.7	Options to deal with <i>presumed anonymous</i> data? .....	103
5.7.1	Potential damage and disadvantage to data subjects .....	106
5.7.2	Consequences of a GDPR violation .....	106
5.7.3	Mandatory damage control when presumed anonymous data is discovered to be personal .....	107
5.7.4	Implementing GDPR requirements for presumed anonymous data .....	112
5.7.5	Summary .....	116
<b>6</b>	<b>Overall summery .....</b>	<b>116</b>

# Towards a Better Understanding of Identification, Pseudonymization, and Anonymization

Bud P. Bruegger <[uld613@datenschutzzentrum.de](mailto:uld613@datenschutzzentrum.de)>

Unabhängiges Landeszentrum für Datenschutz (ULD) Schleswig-Holstein, Germany

Version 1.0

June 2021

## 1 Objective and Approach

When tasked with writing practical guidelines about pseudonymization and anonymization, the author decided that there are still some open questions and that a deeper understanding would be helpful. The present document represents a first attempt at answering open question and fostering a deeper understanding.

The authoritative clarification of these issues initially lies in the hands of the European Data Protection Board (EDPB), after that with courts, and finally with the European Court of Justice (ECJ). This document is thus merely meant as a contribution to the discussions and processes that result in a future authoritative clarification.

In absence of an authoritative position, this document attempts to convince through its internal logic, clarity, and a solid basis in the GDPR. To achieve clarity, emphasis has been put on a precise and mutually consistent use of terminology. It has been attempted to make the basis of reasoning explicit by providing a technical interpretation of concepts related to *identification*. As much as possible, the provided interpretation has been based on the wording of the GDPR and the interpretations have been presented in a hopefully reproducible, comprehensible, and logical manner.

## 2 Outline

Both, *pseudonymization* and *anonymization* are related concepts. In particular, they are both defined in terms of *identification*. For this reason, both topics are discussed in a single section. To gain a precise understanding of the concepts, a first subsection analyses the concrete technical meaning of identification. This is then used to define and understand the two main concepts of this section.

## 3 The concept of identification

The following analyses the concept of identification. For this purpose, it discusses the following topics:

- A definition of *information*, a term used prominently in the GDPR.
- Information elements that are relevant to identification. The precise definition of such data elements is important to provide technical concreteness to the overall discussion.
- A definition of individual-level information and data records. These straightforward definitions facilitate the analysis that follows.
- An analysis of the concept of *linking* information. Linking of information is a key concept for the definition of identification. The distinction of two main types of linking is crucial for the understanding of the overall discussion.
- A definition of identification. Since the GDPR speaks of identified and identifiable persons, the concept of identification is at the very core of understanding pseudonymization and anonymization.
- The risk of identification.
- Measures to reduce the risk of identification.

The in-depth discussion of these concepts, together with precise definitions and concrete examples forms the basis for a clear analysis and understanding of *pseudonymization* and *anonymization*.

### 3.1 Information

A concept that is highly relevant to our discussion is that of *information*. It is central to the definition of both, *personal data* in Article 4(1) GDPR and *pseudonymization* in Article 4(5) GDPR. While different aspects of information are important in different contexts, the following provides a working definition meant to solely support the presented analysis. It thus refrains from attempting to provide a general definition.

Working Definition: ***information***

Information consists of expressions represented either in the form of

- **data**, or
- **knowledge** held by a person.

It also includes ***meta-information*** about data sets, such as information about how these have been created and how the persons described by the data have been selected.

This working definition clarifies, that information consists of more than just technically represented data insofar as it can also exist in “the head” of a person. The knowledge held by persons also illustrates that *information* must be always seen relative to a person, such that typically, different persons are in possession of different information.

Meta-information, particularly about data sets, is highly relevant in the context of identification. Meta-information is often not explicitly expressed in the form of data, but remains in the realm of knowledge held by persons. Its importance in the context of identification shall be illustrated with the following example. Assume that a data set contains a record of a person who is male, blond and has blue eyes. This data by itself does not seem highly identifying. This changes when the meta-information is added that the data set describes school children of some place in India, where blond hair and blue eyes are highly uncommon. In this setting, it may even be possible to uniquely identify the person described by the data.



## 3.2 Information Elements with Relevance to Identification

According to Art. 4(1) GDPR, “[a] natural person [...] can be **identified** [...] by reference to an **identifier** [...] or to one or more **factors specific to the** [...] **identity of that natural person**”. (Highlighting added by the author). This poses the question what exactly is meant by *identifier* and *factor specific to the identity* of a person (in the sequel also called *identity-specific factor*). More generally, which data elements are relevant to identification and how must these legal terms be interpreted technically.

The structure of this part of Art. 4(1) is made explicit in the form of a list: (Structure and italics added by author).

“[A]n 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;”*

So what exactly are *identifiers* and *identity-specific factors*?

Help in the interpretation of these terms comes again from the Article 29 Data Protection Working Party<sup>3</sup>.

They state the following about *identifiers*: “Identification is normally achieved through particular pieces of information which we may call ‘identifiers’ and which hold a particularly privileged and close relationship with the particular individual.” This is used as a basis for the following interpretation:

---

<sup>3</sup> Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data, adopted June 20 2007, WP 136, on page 6, section 3. THIRD ELEMENT: “IDENTIFIED OR IDENTIFIABLE” [NATURAL PERSON], in particular pages 12 through 15.

Interpretation: **identifier**

An identifier is composed of one or several pieces of information that are suitable to identify a person.

The examples of *identifiers* given in Art. 4(1) (see above) provide further insight. They are listed and annotated in the following:

- *Identification numbers and online identifiers*: These are **unique handles** (see definition below).
- *Names*: These are **handles** that are used outside of the **domain** where they are guaranteed to be unique. The Article 29 Data Protection Working Party therefore states<sup>4</sup>: “In order to ascertain this identity, the name of the person sometimes has to be combined with other pieces of information (date of birth, names of the parents, address or a photograph of the face) to prevent confusion between that person and possible namesakes.” On this basis, we propose that a more precise conceptualization is that the name, together with these additional pieces of information, must be considered to be the identifier.
- *Location data*: This kind of information is known to possess a high potential of identification. It comes in very different forms including a point in space (e.g., expressed as latitude and longitude coordinates), or the identifier of a special region (e.g., expressed as a ZIP code or county name).

The Working Party, further cites a commentary by the Commission to clarify that identifiers can either identify a person *directly* or *indirectly*, namely “a person may be identified directly by name or indirectly by a telephone number, a car registration number, a social security number, a passport number or by a combination of significant criteria which allows him to be recognized by narrowing down the group to which he belongs (age, occupation, place of residence, etc.)”.

The examples for identifier provided in Art. 4(1) GDPR hint at a rather loose definition that encompasses many different types of information. The commentary by the Article 29 Data Protection Working Party further widens the definition of *identifier* by adding also unique handles of objects (such as telephone, car, passport) related to a person as well as *combination of significant criteria* (such as age, occupation, place of residence).

Besides the concept of *identifier*, the GDPR also speaks of *factors specific to the identity of a person* (here called *identity specific factors*). The Article 29 Data Protection Working Party provides help for its interpretation. In particular, they state the following: “As regards ‘**indirectly identified or identifiable persons**, this category typically relates to the phenomenon of ‘**unique combinations**’, whether small or large in size. In cases where prima facie the extent of the identifiers available does not allow anyone to single out a particular person, that person might still be ‘identifiable’ because that information combined with other pieces of information (whether the latter is retained by the data controller or not) will allow the individual to be distinguished from others. This is where the Directive comes in with ‘**one or more factors specific to his** physical, physiological, mental, economic, cultural or social **identity**’.”<sup>5</sup> (Highlighting added by the author).

This is used as the basis for the following interpretation:

---

<sup>4</sup> WP 136, page 13, 3<sup>rd</sup> paragraph.

<sup>5</sup> WP 136, page 13, 4<sup>th</sup> paragraph. Note that the Article 29 Data Protection Working Party refers to the Data Protection Directive that pre-dated the GDPR. Their guidance is equally applicable, however, since the wording they interpret is almost (essentially) identical with that of Art. 4(1) of the GDPR.

Interpretation: **identity-specific factor**

An identity-specific factor are a potentially unique combination of information elements that can indirectly identify a person.

The above discussion of the concepts of *identifier* and *identity-specific factor* illustrate that they are difficult to capture precisely since available statements about them leave ample room for interpretation and since they encompass a range of cases and types of information. Furthermore, it may not always be easy to distinguish whether a single or a combination of information elements is an *identifier* or an *identity-specific factor*. For example, in the interpretation of both concepts, the Article 29 Data Protection Working Party speaks of combination: “a combination of significant criteria [...] (age, occupation, place of residence, etc.)” and “unique combinations”, respectively. Also, it remains difficult to understand why a location at which a person is present at a given time should be treated as an *identifier* much rather than as an *identity-specific factor*. What exactly is the difference?

To avoid difficulties with a vague definition of concepts, the following proposes a set of more clearly definable concepts instead of *identifier* and *identity-specific factor*. These are meant to be alternative concepts, not interpretations of the concepts of the GDPR. For this reason, care was taken to avoid a direct conflict of terminology. Most prominently, the term “unique identifier” was avoided as not to imply a relation to the concept of *identifier* used in the GDPR. The concepts proposed are the following:

- **Unique handles,**
- **quasi-identifiers** (including **non-unique handles**), and
- **identity-relevant properties.**

These are defined in the following:

*Unique handles* are often called *unique identifiers* but the term *handle* is used here instead to avoid possible confusion with the use of the term *identifier* in the GDPR.

Definition: **unique handle**

A unique handle is an information element, such as a string or number, with the purpose of referring to a single entity within a pre-defined set of possible entities. Every entity in the set has exactly one handle; the handles of two distinct entities of the set are always different. A unique handle can be seen as an artefact created by an actor as a representation of the *identity* of an entity.

Examples for unique handles include the following:

- First names (given names) given by parents to their children. They are unique in the core family. Should the same first name already be used by other persons in the core family, “tie breakers” such as *junior*, *senior*, *the first*, or *the second* are typically used to render the name unique. Middle names may serve the same purpose.
- Nicknames for people in a group of friends. Nicknames are often used for friends who have the same given name to distinguish them in the group.
- Family names for families living in small communities such as villages where these names were likely unique at the time of assignment.

- Customer numbers assigned by a company to its customers.
- Username or online-identifier.
- E-mail addresses. The assignment of the username component is under the control of the e-mail provider and enforced to be unique. The domain component of the e-mail address then represents the e-mail provider and is guaranteed to be globally unique based on the management of domains by the global organization of the Internet domain name registry. E-mail addresses are thus an example for a globally unique handle.
- Unique handles that represent the identity of devices, such as phone numbers, MAC Addresses, serial numbers, etc.
- Unique handles that represent the identity of vehicles such as license plate numbers or the vehicle identification number.
- A postal address that typically relates to a unique letter box.
- An IBAN or account number of a bank account.

Since unique handles are only unique in a given context, it is practical to define an appropriate term to capture this fact:

Definition: ***identity domain***

An *identity domain* is a context consisting of a group of eligible entities (sometimes called *eligible population*), and an actor (called *domain owner*) who is responsible for issuing *unique handles*, and a procedure to determine the handle of a given entity. Handles in a given identity domain are designed to be unique.

Note that unique handles are sometimes also be used outside of their identity domain. This is for example routinely the case for names (first and family name). When used outside of the domain where they were assigned, they are not guaranteed to be unique any longer. Unique handles use outside of their domain can therefore not be considered unique handles any longer. In many cases, they assume the characteristics of *quasi-identifiers* (see definition below).

Definition: ***non-unique handle***

A *non-unique handle* is an originally *unique handle* that is used outside of its *identity domain* and is therefore no longer guaranteed to be unique. It often has the identification characteristics of a quasi-identifier.

At the example for first/last name pairs, hhainguyen illustrates that *non-unique handles* can still be unique for a significant number of the population. In particular, he created a map showing the number of unique name-pairs in various countries. A static image of his online map<sup>6</sup> is shown in Figure 1. The following table shows the situation in some countries. It was constructed based on extracting precise numbers of unique pairs from hhainguyen's online map<sup>7</sup> and data on the population provided<sup>8</sup> by the United Nations for 2018.

---

<sup>6</sup> See <https://plot.ly/~hhainguyen/74/unique-name-pairs-count-per-country-sourceworldnames-db-and-name-statisticsorg/> (last visited 9/11/2020).

<sup>7</sup> These are accessible via mouse-over pop-up boxes.

<sup>8</sup> As reported by Wikipedia at

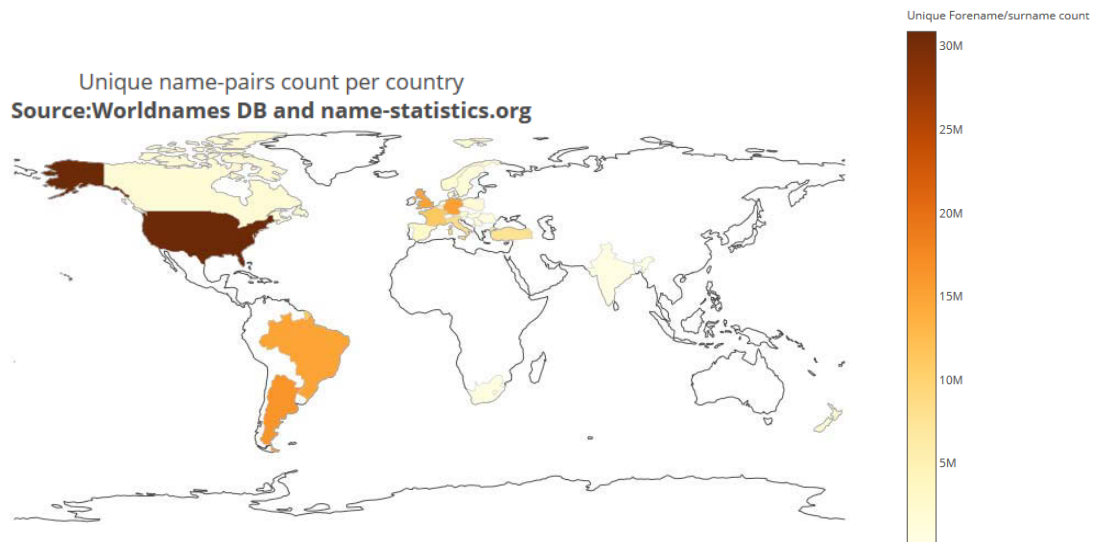


Figure 1: Uniqueness of first/last name pairs according to hhainguyen.

Country	Unique name pairs	Population	Unique per Population
France	11.103 M	64.990 M	17.0 %
Germany	15.370 M	83.124 M	18.8 %
Italy	8.470 M	60.627 M	14.0 %
United Kingdom	14.962 M	67.142 M	22.3 %
United States	30.880 M	327.096 M	9.4 %

Table 1: Unique name pairs per country.

The next concept that is defined is that of quasi-identifiers:

Definition: **quasi-identifier**

A quasi-identifier is composed of one or a combination of information elements that are unique for at least a significant number of persons contained in a data set.

This definition seems to be more or less in line with that given by Wikipedia<sup>9</sup>. The term is extensively used in the context of “anonymization techniques” such as generalization or anatomization (see below). The term is also used by the Article 29 Data Protection Working Party in their *Opinion on Anonymisation Techniques* but without a clear definition.

<sup>9</sup> <https://en.wikipedia.org/wiki/Quasi-identifier>

Typical examples for quasi-identifiers are the following:

- Name, gender, date and place of birth<sup>10</sup>;
- 5-digit ZIP, gender, and date of birth<sup>11</sup>;
- Mobility data<sup>12</sup>;
- Certain kinds of biometrics, such as fingerprints (depending on the size of the candidate population across which it should be close to unique),
- Certain kinds of genetic data, such as DNA (which is unique except in the case of identical twins), or short tandem repeats on the Y chromosome<sup>13</sup>.

Non-unique handles (such as names) are also often part of quasi-identifiers or can in certain contexts even be considered quasi-identifiers by themselves.

The third type of information element is an *identity-relevant property* that is defined in the following:

**Definition: identity-relevant property**

An identity-relevant property is a combination of information elements that has the potential to be unique at least for one or a few persons. This definition is very similar to that of a quasi-identifier. The difference lies in the “power” of identification. In particular, an identity-relevant property may be unique only for rare combinations of values for only one or few persons of a candidate set.

Since unique combinations of values are often unexpected, it is a safe approach to consider any property that is related to a person, the person’s activities and expressions, or any entity closely related to a person as an identity-relevant property. This seems in line with the GDPR’s wording of “factor specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person”. To keep full freedom of definition and avoid any risk that the reasoning would break apart if the term from the GDPR should be interpreted differently, the independent term *identity-relevant property* is used in the sequel.

A simple example that illustrates identity-relevant properties is eye color. It is usually not thought of being identifying, since the common eye colors are shared by large number of persons. However, *red* is one of the possible eye colors<sup>14</sup> and is so rare<sup>15</sup> that it could easily identify a single individual.

---

<sup>10</sup> This combination is for example used in some national unique schemes for unique handles such as the Italian tax number.

<sup>11</sup> See for example: L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000, <https://dataprivacylab.org/projects/identifiability/paper1.pdf> (last visited 5/11/2020).

<sup>12</sup> See for example: de Montjoye, Y., Hidalgo, C., Verleysen, M. et al. Unique in the Crowd: The privacy bounds of human mobility. *Sci Rep* 3, 1376 (2013). <https://doi.org/10.1038/srep01376>

<sup>13</sup> See Melissa Gymrek; Amy L. McGuire; David Golan; Eran Halperin; Yaniv Erlich (18 January 2013), "Identifying personal genomes by surname inference", *Science*, 339 (6117), Bibcode:2013Sci...339..321G, doi:10.1126/SCIENCE.1229566, PMID 23329047, Wikidata Q29619963.

<sup>14</sup> See for example, Rebecca E., Rare Human Eye Colors, Sciencing, Updated July 20, 2018, <https://sciencing.com/rare-human-eye-colors-6388814.html> (last visited 10/11/2020).

<sup>15</sup> Red eyes seem to be related to albinism and Wikipedia states that in Europe and the United States, the prevalence of albinism is about 1 in 20'000 (see [https://en.wikipedia.org/wiki/Albinism\\_in\\_humans#Epidemiology](https://en.wikipedia.org/wiki/Albinism_in_humans#Epidemiology), last visited 10/11/2020).

While red eye color is very rare worldwide, other properties may be very rare in certain countries or regions. For example, Jewish confession is rather rare in Iran or blond hair is rare in certain Asian countries.

While in these simple examples, the rareness may initially be unexpected, it then becomes rather evident. In contrast, rareness may often be more difficult to recognize and understand in larger and more complex combinations of information elements.

Unique combinations can also be present in little structured data sets. A well-known example for this is the “anonymized” search history published by AOL. Based among others on place and family names contained in the searches of an initially pseudonymous user (AOL Searcher No. 4417749), the person behind it could be re-identified<sup>16</sup>.

### 3.2.1 Uniqueness and the Dimension of the Data Set

Uniqueness seems to be very common in so-called *high-dimensional data sets*<sup>17</sup>.

Definition: **dimension** of a data set

The *dimension* of an individual-level data set is simply the number of attributes that it contains for each person. In a tabular representation of the data set, it corresponds to the number of columns (where rows are data records linked to a single individual).

In high-dimensional data sets, every attribute in the data set is considered to be a dimension of its own. For every dimension, an axis can be imagined. Attribute values can then be seen as coordinates along their axes. Every actual data record (that is composed of a tuple of attribute values) can then be seen as a point in this multi-dimensional space.

In this setting, the uniqueness of a data record can be understood as the distance between the data record (as a point in space) to all the closest data records (i.e., points) near by. If a data point is far from all other data points, it is rather unique; if it is part of a cluster of points that are mutually close, it is far less unique. Obviously, the more unique a data record is, the more potential it has to identify a data subject.

In this context, it has been argued that the higher the dimension of a data set, i.e., the more attributes it contains, the more likely it is that at least some data records are highly unique. The reasoning behind this is that when a data record is close to others looking only at a subset of attributes, it is likely to distinguish itself from these records in the other attributes. This pattern becomes more likely with increasing dimension of the data set. In other words, finding points that are close when considering all attributes becomes less likely with increasing number of attributes.

---

<sup>16</sup> See Michael Barbaro and Tom Zeller Jr., A Face Is Exposed for AOL Searcher No. 4417749, New York Times, August 10, 2006, [https://archive.nytimes.com/www.nytimes.com/learning/teachers/featured\\_articles/20060810thursday.html](https://archive.nytimes.com/www.nytimes.com/learning/teachers/featured_articles/20060810thursday.html) (last visited 10/11/2020).

<sup>17</sup> See for example, Aggarwal, Charu C. (2005). "On k-Anonymity and the Curse of Dimensionality". VLDB '05 – Proceedings of the 31st International Conference on Very large Data Bases. Trondheim, Norway. CiteSeerX 10.1.1.60.3155. ISBN 1-59593-154-6, <http://www.charuaggarwal.net/privh.pdf> (last visited 10/11/2020).

The Article 29 Data Protection Working Party emphasizes the identification potential of high-dimensional data in their *Opinion on Anonymisation Techniques*<sup>18</sup>. It also provides an example where the identification of data subjects was possible due to the uniqueness of data records in a high-dimensional data set. Namely, this is the well-publicized identification of persons in the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix<sup>19</sup> that was linked against the Internet Movie Database. The Article 29 Data Protection Working Party describes the situation as follows<sup>20</sup>:

- “In other words, the [...] selection of just 8 rated movies constituted a *fingerprint* of the expressed ratings, *not shared between two data subjects within the database.*”, and
- “Even the injection of noise fails to bring records *sufficiently close together* to share that same *multi-dimensional region.*”

### 3.3 Individual-Level Information and Data Records

The following provides definitions for *individual-level information* and *data records*. These are helpful for the further discussion.

Definition: ***individual-level information***

Individual-level information is information, where information elements can be attributed to a single person (i.e., an individual). In statistics, this is often called *micro data*.

All information elements with relevance to identification are individual-level information elements.

Individual-level information contrasts with group-level information where information elements are attributed to groups of persons. Prime example for group-level information comes from statistics where the properties of a group, such as the male or the female population or age groups, are described by statistical values such as average, median, minimum or maximum.

Definition: ***data record***

A data record is a subset of a data set that contains all information elements related to a single person.

The concept of data record is well-established and introduced here to support the definition of linking below.

### 3.4 Linking of Information

The linking of information lies at the core of the definition of *identification* that will be given in the next section. The precise meaning of linking is therefore discussed here. The scope of the discussion is limited to individual-level information.

---

<sup>18</sup> See page 30 of Article 29 Data Protection Working Party, WP216, Opinion 05/2014 on Anonymisation Techniques, Adopted on 10 April 2014, [https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216\\_en.pdf](https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf) (last visited 15/12/2020).

<sup>19</sup> Arvind Narayanan, Vitaly Shmatikov: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008:111-125, <https://doi.org/10.1109/SP.2008.33>, [https://www.cs.utexas.edu/~shmat/shmat\\_oak08netflix.pdf](https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf) (last visited 15/12/2020).

<sup>20</sup> See footnote 18, page 30. Highlighting added by the author.



Ideally, linking permits to combine two data sets such that the respective data records belonging to the same person are identified and combined. Since this is not always achievable, the following definition is formulated a little more general. Different types of linking will then be discussed after the general definition.

Definition: **Linking**

The objective of linking is to obtain information about how data records (or single attributes) of one data set or information collection relate to the data records of another one.

There are two main mechanisms of linking, usually called *deterministic linking* and *probabilistic linking*.<sup>21</sup> Additional literature about linking of data sets were provided for example by Leicester Gill<sup>22</sup> and Statistics New Zealand<sup>23</sup>.

To prepare the distinction of different types of linking, two different types of value scales are defined here:

Definition: **discrete value**

A discrete value is expressed on a scale that is based on a pre-defined set of possible values. Examples for discrete values are nominal values (such as names, strings, or colors) and integer numbers (such as a year). Discrete values can be compared by checking on equality.

Definition: **continuous value**

A continuous value is expressed on a scale on which there exists an infinite number of values between any two values. Continuous values are for example measurements expressed on a ratio scale or as real (floating point) numbers (such as blood pressure or weight). The comparison of continuous values is based on the notion of difference<sup>24</sup>. When continuous values are the result of measurement or observation, they are typically subject to limited precision, accuracy, and random errors. The concept of equality of two continuous values therefore does not exist. Much rather, continuous values can be similar, close, or correlated.

---

<sup>21</sup> See for example, Australian Government, Open Data Toolkit, Data Linking, <https://toolkit.data.gov.au/Data Linking Information Series Contents page.html>

<sup>22</sup> Leicester Gill, 2001, Methods for Automatic Record Matching and Linkage and Their Use in National Statistics, Issue 25 of National statistics methodology series, Great Britain Office for National Statistics, ISBN 1857744209, 9781857744200, <https://webarchive.nationalarchives.gov.uk/20160107223300/http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/gss-methodology-series--25--methods-for-automatic-record-matching-and-linkage-and-their-use-in-national-statistics.pdf>, last visited 15/10/2020.

<sup>23</sup> Statistics New Zealand, 2006, Data Integration Manual, 2nd edition, ISBN 0-478-26971-4, <http://archive.stats.govt.nz/methods/data-integration/data-integration-manual-2edn>, last visited 15/10/2020.

<sup>24</sup> The difference is usually defined in terms of a distance function.

Based on this distinction of values, two types of linking can be distinguished. The first kind of linking is based on equality of discrete values:

Definition: **Deterministic Linking**

Deterministic linking establishes relationships between data records of distinct data sets based on the comparison of *discrete* information elements for **equality**.

Deterministic linking is illustrated in Figure 2:

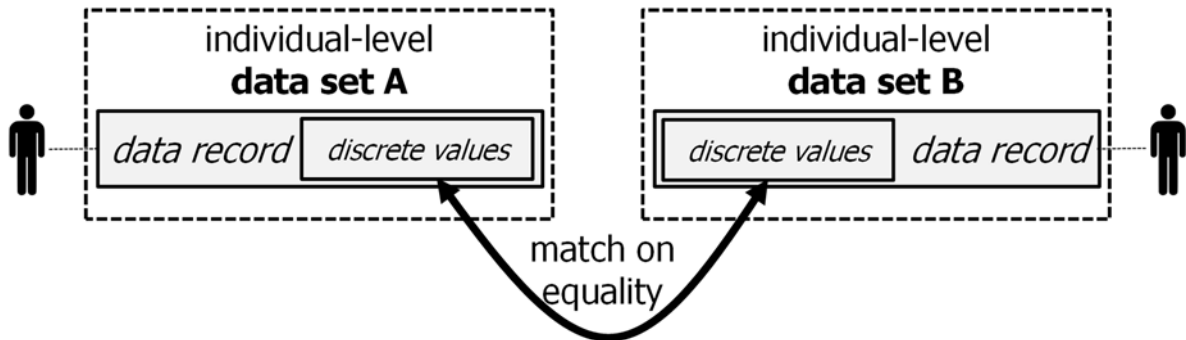


Figure 2: Deterministic linking.

When the discrete values that are compared act as identifiers for the person, the matches are expected to be **unique**, i.e., a data record in one data set matches exactly one data record in the other.

When the discrete values do not uniquely identify individuals, matching may be **ambiguous**. In this case, a data record of one data set may match several data records in the other data set (and vice-versa). Assuming in both data sets, distinct data records belong to distinct persons, such ambiguity introduces uncertainty: instead of finding the matching person in the other data set, a possibly small set of “candidates” is found. Often, such uncertainty can be removed or further reduced in additional steps by matching with additional data sets.

The second kind of linking is based on the similarity, proximity, or correlation of continuous values:

Definition: **Probabilistic Linking**

Probabilistic linking establishes relationships between data records of distinct data sets based on the comparison of continuous values for **similarity, proximity, or correlation**.

Probabilistic linking is illustrated in Figure 3:

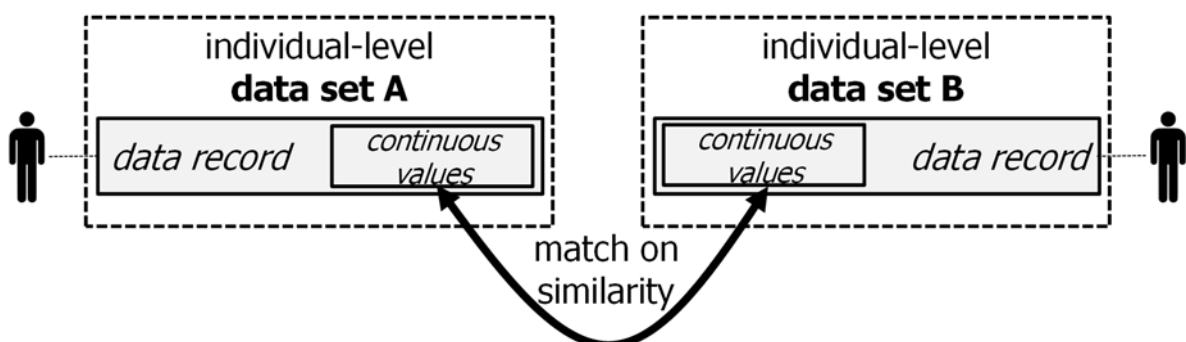


Figure 3: Probabilistic linking.

Probabilistic linking is typically based on continuously valued quasi-identifiers or identity-relevant properties. A precise match with equal values in both data sets is highly unlikely. Therefore, only a closeness, similarity, or correlation of the values can be determined. The resulting relation between data records in the different data sets is therefore not Boolean (i.e., “belong to the same person”, “belong to different persons”). Much rather, the relation expresses a probability that the data records actually belong to the same person.

Probabilistic linking is sometimes also used to compare discrete values that contain errors. A typical application is the comparison of names (i.e., strings) that may contain spelling errors. Instead of comparing for equality, the strings are then often compared for similarity. For this purpose, a string distance function<sup>25</sup> can be chosen to quantify the similarity of two strings. The probably best-known string distance functions are the Hamming distance<sup>26</sup> and the Levenshtein distance<sup>27</sup>. It measures how many characters need to be changed to arrive from one string to the other. Note that beyond simply comparing the similarity of strings, there are also highly efficient methods of “fuzzy matching”<sup>28</sup>. An example of a sophisticated and multi-faceted way of matching is provided by recent work by Oana Goga et al.<sup>29</sup>. Another sophisticated approach of “fuzzy matching” is described by Ranjan Kant and Piyush Sagar Mishra<sup>30</sup>. The latter links entities based on clustering of data as described by McInnes et al<sup>31</sup>.

A good example that helps to understand probabilistic linking is biometric matching. Assume that two biometric data records (e.g., corresponding to a fingerprint) are compared to determine whether they belong to the same person.

This is illustrated with the visualization by Dhir et al<sup>32</sup> in Figure 3.

Figure 4a shows the probability distribution of biometric data observed from a “genuine” person and that of a potential imposter. It shows how a threshold for the match score needs to be chosen in order to decide whether the observed biometric data belongs to the genuine person or an imposter.

---

<sup>25</sup> Wikipedia lists some string distance functions at [https://en.wikipedia.org/wiki/String\\_metric](https://en.wikipedia.org/wiki/String_metric) (last visited 11/11/2020).

<sup>26</sup> See for example Wikipedia at [https://en.wikipedia.org/wiki/Hamming\\_distance](https://en.wikipedia.org/wiki/Hamming_distance) (last visited 11/11/2020).

<sup>27</sup> See for example Wikipedia at [https://en.wikipedia.org/wiki/Levenshtein\\_distance](https://en.wikipedia.org/wiki/Levenshtein_distance) (last visited 17/11/2020).

<sup>28</sup> See for example, Van den Blog, Super Fast String Matching in Python, October 14, 2017, <https://bergvca.github.io/2017/10/14/super-fast-string-matching.html> (last visited 17/11/2020), the python project TD-IDF matcher at <https://pypi.org/project/tfidf-matcher/> (last visited 17/11/2020) and the blog post by Josh Taylor, Fuzzy matching at scale, July 2 2019, <https://towardsdatascience.com/fuzzy-matching-at-scale-84f2bfd0c536> (last visited 17/11/2020).

<sup>29</sup> Goga, Oana & Loiseau, Patrick & Sommer, Robin & Teixeira, Renata & Gummadi, Krishna P.. (2015). On the Reliability of Profile Matching Across Large Online Social Networks. 10.1145/2783258.2788601. <https://arxiv.org/abs/1506.02289> (last visited 17/11/2020)

<sup>30</sup> Ranjan Kant and Piyush Sagar Mishra, An Ensemble Approach to Large-Scale Fuzzy Name Matching, March 28, 2019, <https://medium.com/bcggamma/an-ensemble-approach-to-large-scale-fuzzy-name-matching-b3e3fa124e3c> (last visited 17/11/2020).

<sup>31</sup> L. McInnes, J. Healy, S. Astels, *hdbscan: Hierarchical density based clustering* In: Journal of Open Source Software, The Open Journal, volume 2, number 11. 2017, as implemented in <https://github.com/scikit-learn-contrib/hdbscan> (last visited 17/11/2020).

<sup>32</sup> The figure is copied from: DHIR, & VIJAY, & AMARPREET, SINGH & RAKESH, KUMAR & GURPREET, SINGH. (2010). BIOMETRIC RECOGNITION: A MODERN ERA FOR SECURITY. International Journal of Engineering Science and Technology. 2. [https://www.researchgate.net/publication/50315614\\_BIOMETRIC\\_RECOGNITION\\_A\\_MODERN\\_ERA\\_FOR\\_SECURITY](https://www.researchgate.net/publication/50315614_BIOMETRIC_RECOGNITION_A_MODERN_ERA_FOR_SECURITY), last visited 15/10/2020. Figure available under Creative Commons Attribution 4.0 International License.

The Figure illustrates that any choice of threshold leads to two types of false decisions:

- The decision that the biometrics of the genuine person does not belong to the genuine person (aka. *false negative* or **false non match**); and
- The decision that the biometrics of the imposter belongs to the genuine person (aka. *false positive* or **false match**).

Figure 4b illustrates how different choices of the threshold lead to different probabilities for false negatives and false positives. It further shows how different purposes (forensic applications, civilian applications, and high security applications) tend to make a different choice for the threshold.

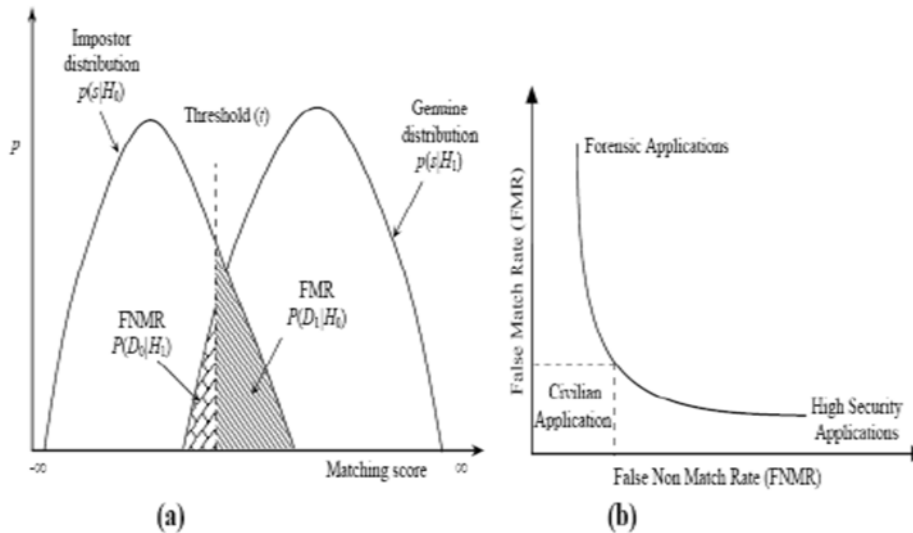


Figure 4: Biometric matching (by Dhir et al).

### 3.4.1 Special kinds of Linking

Linking most commonly establishes relationships between two complete data records based on a comparison of their data values. There are some special cases to keep in mind when reasoning about linking. These are described in the following.

#### Linking based on implicit order:

While obvious, it is sometimes overseen that linking may not only be based on the comparison of data values that are explicitly contained in a data set, but also on the implicit order of data records in a data set. This is for example possible when two “de-identified”<sup>33</sup> data sets are derived from the same source. Neither of the two data sets contain any data elements that would allow linking based on comparison. But if the (implicit) order of the records is known to be the same, a linking of records belonging to the same person is anyhow possible.

#### Model-based linking and inference:

The most straight forward form of linking is possible when the two data sets that are compared contain some common attributes, such as handles or quasi-identifiers. Comparison is also possible

<sup>33</sup> The term “de-identified” is not defined here and is therefore put inside quotation marks.

when the data sets contain no common attributes. In this case, a functional or probabilistic model can be used that establishes the relationship between the attributes contained in one data set with those contained in the other. Examples include the following:

- Inference of a person's gender based on certain diseases (e.g., breast and prostate cancer);
- Inference of trip destinations from speed and time data<sup>34 35</sup>;
- Inference of location tracks base on smartphone sensors<sup>36</sup>;
- Inference of political orientation from profile facial images<sup>37</sup>.

It is expected that with the increasing use of artificial intelligence and machine learning based on very large data sets, the potential sophistication and use of model-based linking will significantly increase. An overview of approaches is given by Asher et al<sup>38 39</sup>. The constant technical advances in artificial intelligence<sup>40</sup> may significantly increase the importance of model-based linking.

Another example, how linking can be more complex than comparing attributes across two data sets is the fact that it is possible to extract personal data used during training from neural networks. In particular, neural networks can "memorize" training data and render it possible to extract them. Further detail is provided in two articles by Carlini et al<sup>41,42</sup>.

---

<sup>34</sup> Rinku Dewri, Prasad Annadata, Wisam Eltarjaman, and Ramakrishna Thurimella. 2013. Inferring trip destinations from driving habits data. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society* (WPES '13). Association for Computing Machinery, New York, NY, USA, 267–272. DOI:<https://doi.org/10.1145/2517840.2517871>

<sup>35</sup> Xianyi Gao, Bernhard Firner, Shridatt Sugrim, Victor Kaiser-Pendergrast, Yulong Yang, and Janne Lindqvist. 2014. Elastic pathing: your speed is enough to track you. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 975–986. DOI:<https://doi.org/10.1145/2632048.2632077>

<sup>36</sup> S. Saha, S. Chatterjee, A. K. Gupta, I. Bhattacharya and T. Mondal, "TrackMe – a low power location tracking system using smart phone sensors," 2015 International Conference on Computing and Network Communications (CoCoNet), Trivandrum, 2015, pp. 457-464, doi: 10.1109/CoCoNet.2015.7411226.

<sup>37</sup> Kosinski, M. Facial recognition technology can expose political orientation from naturalistic facial images. *Sci Rep* 11, 100 (2021). <https://doi.org/10.1038/s41598-020-79310-1> (last visited 15/1/2021).

<sup>38</sup> Mudgal, Sidharth & Li, Han & Rekatsinas, Theodoros & Doan, AnHai & Park, Youngchoon & Krishnan, Ganesh & Deep, Rohit & Arcaute, Esteban & Raghavendra, Vijay. (2018). Deep Learning for Entity Matching: A Design Space Exploration. SIGMOD '18: Proceedings of the 2018 International Conference on Management of Data. 19-34. 10.1145/3183713.3196926.

<sup>39</sup> Asher, Jana & Resnick, Dean & Brite, Jennifer & Brackbill, Robert & Cone, James. (2020). An Introduction to Probabilistic Record Linkage with a Focus on Linkage Processing for WTC Registries. *International Journal of Environmental Research and Public Health*. 17. 6937. 10.3390/ijerph17186937.

<sup>40</sup> See for example William Fedus, Barret Zoph and Noam Shazeer, Switch Transformers: Scaling to Trillion Parameter Models with simple and efficient Sparsity, <https://arxiv.org/pdf/2101.03961.pdf> (last visited 15/1/2021).

<sup>41</sup> Extracting Training Data from Large Language Models, Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, Colin Raffel, <https://arxiv.org/abs/1802.08232> (last visited 8/1/2021)

<sup>42</sup> Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium (SEC'19)*. USENIX Association, USA, 267–284, <https://arxiv.org/abs/2012.07805> (last visited 8/1/2021).

### Attribute Linking:

Most often, linking is concerned with relating two complete records that belong to the same person or, more generally entity. Such linking is therefore also often called *record linkage* or *entity resolution*<sup>43</sup>.

There are cases, however, where it is not possible to link two complete records, but still to link a subset of attributes from one record to the other. This is called attribute linking<sup>44</sup>. The following example shall illustrate this.

Assume that an initial data set contains full names, addresses and some additional attributes of persons. To prevent linking, this data set is modified such that the names are dropped and the addresses are reduced to just the ZIP code of the original address. After ascertaining, that a minimal number  $k$ <sup>45</sup> of persons falls in every ZIP code area, the data is then published.

Evidently, the fact that every ZIP code area contains the records of at least  $k$  persons renders deterministic linking highly ambiguous. When one of the attributes has the same value for all persons in a ZIP code area, however, it is indeed possible to learn this one attribute value from the published data set and link it to all persons residing at this ZIP code<sup>46</sup>. This is also called a *homogeneity attack*.

Such attribute linking which associates a single attribute value on one side with data records on the other side is also considered to be linking. Attribute linking can also enable identification of a person.

## 3.5 Identification

Based on the definition of the concepts above, the present section provides a technical interpretation of the meaning of *identification*. The latter is a key concept used in the GDPR. In particular, it is central to the definition of *personal data* in Art. 4(1), the concepts of *pseudonymization* in Art. 4(5), and that of *anonymous data* in Recital 26.

The present section analyses in particular the technical concepts behind the terms *identified* and *identifiable* (see Art. 4(1) GDPR). It also explains the difference between *direct* and *indirect* identification (see also Art. 4(1) GDPR).

The following analyses some wording of the GDPR to draw some conclusions about the concept of *identification*.

As evident in Art. 4(1) GDPR, **identification is defined relative to a data set**: “[A]n identifiable natural **person** is one who can be *identified*, directly or indirectly, in particular *by reference to an identifier [...] or to one or more factors specific to the [...] identity of that natural person;*” (Highlighting

---

<sup>43</sup> See for example, [https://en.wikipedia.org/wiki/Record\\_linkage](https://en.wikipedia.org/wiki/Record_linkage) (last visited 19/11/2020).

<sup>44</sup> See for example Section 2.2, Benjamin C.M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. 2010. Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques (1st. ed.). Chapman & Hall/CRC, DOI: 10.1201/9781420091502, <http://www.gbv.de/dms/tib-ub-hannover/630276005.pdf> (last visited 16/12/2020).

<sup>45</sup> This example is based on the concept of  $k$ -anonymity as described in Samarati, P. and L. Sweeney. “Protecting privacy when disclosing information:  $k$ -anonymity and its enforcement through generalization and suppression.” (1998).

<sup>46</sup> This “homogeneity attack” has first been described by Machanavajjhala et al. and is the basis for the definition of  $l$ -diversity as describe in Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007.  $L$ -diversity: Privacy beyond  $k$ -anonymity. ACM Trans. Knowl. Discov. Data 1, 1 (March 2007), 3–es. DOI:<https://doi.org/10.1145/1217299.1217302>.

with italic and snipping added by the author). Clearly, the concept of identification is related to a data set that contains identifiers or identity-specific factors.

The wording of Art. 4(1) GDPR further implies that **identification is performed through linking**: Establishing a “reference to” a data element such as an identifier or an identity-specific factor is nothing other than linking.

It is evident from Art. 11(2) GDPR, that **identification requires an actor** who performs the identification: “Where [...] the controller is able to demonstrate that it is not in a position to identify the data subject [...]”. Here, the controller is stated to be the actor performing the identification. The wording implies, that there may be other actors (than the controller) who may have different capabilities for identifying the data subject: The fact that the controller is unable to identify the data subject does not exclude that other actors are able to do it. Whether a data subject is identified or identifiable can thus be answered only relative to a given actor. A universal concept of *identified* therefore only exists in the sense that in some cases, any possible actor may be able to identify the data subject.

According to Art. 4(1) GDPR, **identification** can be **direct** or **indirect**: “[A]n identifiable natural **person** is one who can be *identified*, directly or indirectly [...]”. Art. 4(5) GDPR provides further insight in the meaning of direct and indirect by using the following wording: “personal data [that] can no longer be attributed to a specific data subject without the use of **additional information**”. Here, we interpret the concept of “attribution to a specific data without the use of additional information” as direct identification and “attribution to a specific data with the use of additional information” as indirect identification.

The elements that arise from the above analysis of the GDPR are arranged visually in Figure 5. Namely, the figure shows that identification is performed by an *actor* who identifies, that identification refers to a *data set* that is represented by a single *data record*, and that identification can be *direct* or *indirect*, depending on whether *additional information* is used.

Possible **actors** that are relevant to the discussion are the controller and possible processors (i.e., organizations), persons acting under the authority of the controller or of the processor (i.e., employees and thus natural persons), third-party recipients of data (see Art. 4(9) and (10) GDPR), and any other thinkable actor (including attackers) who gain access to the data record.

The figure further shows the assets that are readily available to the actor who performs the identification. This includes *knowledge* of the actor, *data* that the actor owns or otherwise has ready access to, as well as *methods of interaction* that the actor can use to interact with the data subject. Examples for interaction include communication systems such as telephone or e-mail, shipping systems that allow the actor to send physical objects to the data subject, or methods that enable the physical meeting of the actor with the data subject.

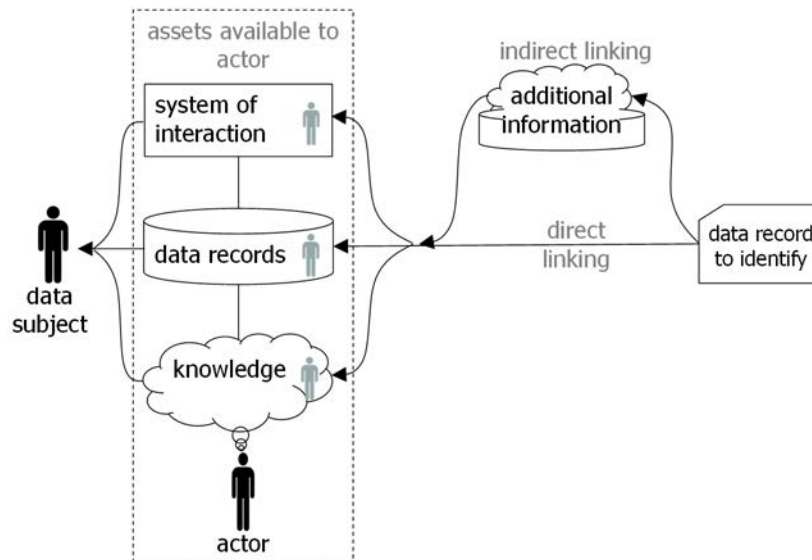


Figure 5: Identification of a data subject.

Based on this figure, the precise meaning of *identified* and *identifiable* are now discussed.

The following informal definition seems straight forward: A natural person is considered to be identified when it becomes clear to the actor that it is indeed this person who is described by a data record at hand.

But what exactly does this mean in a concrete technical setting? The question is approached by using the specification of the technical setting of Figure 5 above, the interpretation provided by the Article 29 Data Protection Working Party, and some generalization.

The Article 29 Data Protection Working Party states that “Concerning ‘directly’ identified or identifiable persons, the name of the person is indeed the most common identifier, and, in practice, the notion of ‘identified person’ implies most often a reference to the person’s name.”

To better understand this statement, the following question arises: What is the role of persons’ names here?

The most likely role of **names** is that they are the primary handles used for persons used in the knowledge (held by actors who are persons). Based on this thought, the above statement could be paraphrased as follows: “the notion of ‘identified person’ implies most often a reference to the **primary person handle** used by the actor.”

Since the Article 29 Data Protection Working Party indicates that it is not always a reference to the name, the question arises which other possibilities there are.

Looking at the figure above, it is evident that the information accessible to actors is not limited to knowledge, but includes also data. Since names are not guaranteed to be unique, they usually are not used as primary handles for persons in data sets. This leads to the thought that a person can also be identified by a reference to a **unique handle used in the available data sets**. This option seems particularly relevant when the actor who identifies is an organization, much rather than a natural person.

The following examples shall further support this interpretation: A tax authority considers a person to be identified when it knows the person’s tax identifier. Similarly, a company considers a customer to be identified when it can associate the person with a customer number. Help desks of the actors are very helpful when trying to understand what *identified* means, since they typically start the interaction with a data subject with questions that lead to an identification.

In the examples above, the tax authority and of the company associate the data subject with a virtual person present in their virtual (electronic) model of the world; in the example of human actors who



identify data subjects by name, the association is to a person in the mental model of the world. The association of the data subject with entities in these models enables the actors to operate the model relative to that data subject. But is identification only concerned with models of the world?

While a significant number of processing operations may primarily be concerned with models of the world, some are also concerned with interaction with the real world. This is modelled in the above figure as interactions with the data subject. Here, a data subject is considered to be identified when the information about how to interact with it is available. In communication systems, this is typically some kind of address (such as an e-mail address or telephone number); for financial transactions, this may be an IBAN or bank account number; for shipping, this is typically a shipping address; for physical interaction, it is the information necessary to physically meet the data subject. Addresses and information about meeting a person can be considered to be unique handles in their own right.

This understanding of *identified* is now integrated with the technical setting of Figure 5. Here, the starting or end point of the identification is a **data record** belonging to the data subject and the way to establish relations between this data record and the assets available to the actor is **linking**. This is used to yield the following definition:

Definition: ***identified***

A data subject described by a *data record* is considered to be *identified* when a whole data record, a subset thereof, or data elements that are derived from it can be *linked* to a *unique handle* for persons used

- in a model of the world (i.e., knowledge) by a human actor,
- in a virtual model of the world (i.e., data) available to the actor,
- as address in some real-world interaction system accessible to the actor.

The linking can be deterministic or probabilistic. For a data subject to be *identified*, deterministic linking needs to be unique and probabilistic linking must single out exactly one person with sufficiently high probability.

The direction in which identification is achieved is irrelevant: Either identification yields the person described by a given set of data elements, or it yields the data elements belonging to a given person.

The linking can be based on the comparison of unique handles, quasi-identifiers, identity-relevant properties, or combinations thereof. It results in the association between the initial data record and a mental representation, data record, or interaction address of the related person in the domain of the actor.

As evident from the figure, the linking can be performed *directly* from the initial data set to information available to the actor, or *indirectly* via first linking to additional information that is then in turn linked to the information available to the actor. This is captured in the following definitions:

Definition: ***direct identification***

Direct identification is based on linking between the initial data record and information available to the actor without the use of additional information.

Definition: ***indirect identification***

Indirect identification is based on linking between the initial data record with additional information and the linking of this additional information with information available to the actor.

Now that identified has been defined, it is possible to address identifiable.

Recital 26 GDPR states the following<sup>47</sup>: “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.”

This is reworded in the following definition:

**Definition: *identifiable***

A data subject described by a *data record* is considered to be *identifiable* if any actor exists at present or in the future who is able to identify (i.e., render identified) the data subject by using any realistically available additional information and linking methodology<sup>48</sup>.

Note that the concept of *identifiable* is not easy to evaluate since the evaluator may not know about all possible actors and the additional information and linking methodology available to them. In addition, such actors, additional information, and linking methodology may not yet exist at the present time but only materialize in the future.

### 3.6 Risk of identification

Considering a data record pertaining to an identifiable data subject, an important question is how high the risk is that an actor can indeed identify the person behind the data record.

To inquire this further, the following annotates the technical setting of Figure 5 with factors that influence the ability to identify the person behind the data record. The result is illustrated in Figure 6 that is discussed in the sequel.

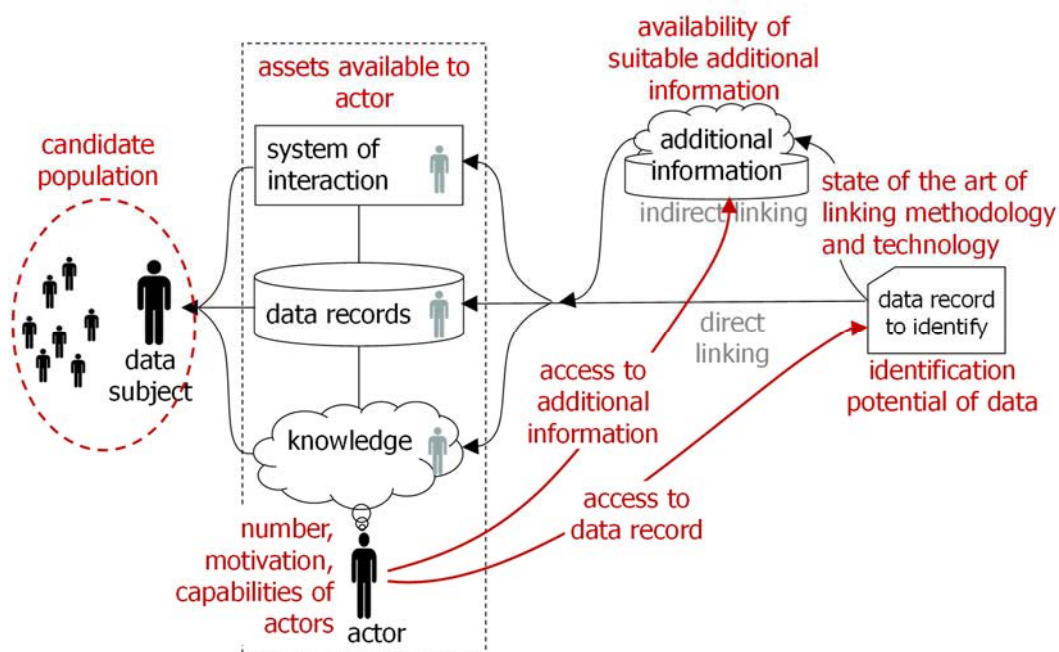


Figure 6: Factors that influence the risk of identification by a given actor.

<sup>47</sup> In the 3<sup>rd</sup> sentence.

<sup>48</sup> This definition is aimed to be in line with the sentence 3 and 4 of Recital 26 GDPR.

- **Candidate population<sup>49</sup>:** The natural person to be identified is always part of a candidate set. If nothing is known, this is the whole world population. But in most cases, the context in which the data were used, some meta data, or some prior, non-conclusive identification attempts can significantly limit the set of candidates. It is now evident that the smaller the candidate population is, the higher the likelihood of a successful identification.
- **Number of actors with access to the data record:** Evidently, actors can only identify a data record if they have access to it. This can be used to control the number of actors who can potentially identify the person behind a data record. For example, if the data is published, any possible actor can attempt an identification; if the access to the data record is controlled, only the actors who have access to the data can attempt identification. It is evident that reducing the number of actors who can attempt identification also reduces the likelihood of identification.
- **Motivations and capabilities of actors:** It is evident that the likelihood of identification depends on how motivated the various possible actors are and what capabilities they dispose of. Actors who are not motivated, will not bother to even try to identify the person; actors who see a high value in the identification and possess the capability of conducting possibly complex and costly linking operations are much more likely to identify the person behind the data record.
- **Overlap of actor's information with candidate population:** The more overlap between the population that an actor manages in its data or knowledge, the higher the likelihood of identification. For example, a social media operator is more likely to identify its users than it is to identify persons who are not active on its platform<sup>50</sup>. Similarly, employees accessing the data under the authority of the controller are more likely to identify a data subject when it is among their friends, neighbors, or other kinds of acquaintances.
- **Types and richness of information assets available to the actor:** Since identification is based on linking, the data available to the actor must have a thematic overlap with the data record to be identified. Identification is therefore more likely, the more the thematic overlap is. Actors with rich data sets that cover a wide range of thematic areas related to persons are more likely to successfully identify data subjects.
- **General availability of additional information:** It is intuitively clear that indirect identification becomes the easier, the more additional information is generally available. The availability has experienced a drastic increase in the recent years with digitalization entering ever more aspects of our lives, more and more transactions being conducted in virtual worlds, sensors becoming ever more affordable, powerful, and ubiquitous, and vehicles, personal devices, and IoT becoming important collectors of personal data. Technological progress and potential benefits of big data to society further fuel this trend. The only possible limiting factors may be introduced by legislators and policy makers.
- **Access to additional information:** When the identification is indirect and thus requires additional information, access to suitable additional information is a prerequisite for identification. Even when information is public, the actor has to find the relevant information. In some cases, this can require a significant effort to weed through large

---

<sup>49</sup> Note that this seems closely related to the concept of *identifiability set* proposed on page 30 in Andreas Pfitzmann and Marit Hansen, A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, Version v0.34, Aug. 10, 2010, [http://dud.inf.tu-dresden.de/literatur/Anon\\_Terminology\\_v0.34.pdf](http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf) (last visited 19/1/2021).

<sup>50</sup> At least this holds under the assumption that the social media operator does not collect data about non-users from other sources.

amounts of data. It may also require specific capabilities to recognize data as relevant to the identification or to bring it into a format<sup>51</sup> that is suitable for linking. When additional information is not public and held by a third party, special legal powers may be necessary to access the data. For example, based on a warrant, law enforcement agencies may gain access to of an ISP about the assignment of IP Addresses to persons.

- **Linking methodology and computing power accessible to the actor:** Active research constantly pushes the frontiers of the know-how of how to identify, re-identify, or de-anonymize data. The possibility to use artificial intelligence to build models that are trained by massive amounts of data to permit the correlation of linking of data sets further emphasizes this development. An actor’s capability to use state of the art identification techniques and have access to the necessary computing power may well determine the likelihood of identification.
- **Identification potential of the data record itself:** Most discussions of pseudonymization and anonymization focus predominantly or even exclusively on this factor. Evidently, the data content determines the ease with which a data record can be linked to the information assets in possession of an actor. For example, the presence of widely used unique handles renders such linking and thus identification relatively easy. When such handles are absent, and linking has to be performed based on identity-relevant properties, identification may only succeed in rare cases where the properties represent rare combinations of values and thus can single out a person.

The factors that influence the risk of identification were explicitly listed as a basis for the discussion of how the risk of identification can be reduced (as is the objective to both, pseudonymization and anonymization). In an attempt to obtain a comprehensive list of such factors, the above analysis was guided by the technical model of identification given in Figure 5. It is hoped that this approach leads to a deeper understanding of risk reduction than would be possible when looking at only the most frequently discussed last factor of the above list.

In a blog post<sup>52</sup>, Paul Francis expresses a very similar model of the risk of identification. He illustrated it in the following figure copied from the blog post. It evidently confirms a subset of the above mentioned risk factors—at times with a different terminology.

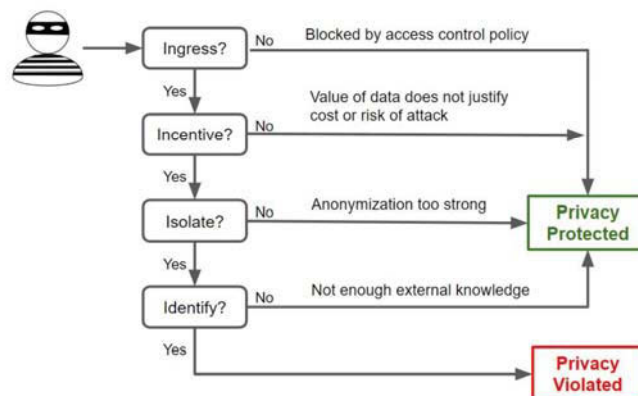


Figure 7: Risks of identification according to P. Francis.

<sup>51</sup> This may for example require to parse and reformat data or to translate information from other languages.

<sup>52</sup> Paul Francis, The five private eyes, Part 1: The surprising strength of de-identified data, June 24, 2020, aircloak, <https://aircloak.com/the-five-private-eyes-part-1-the-surprising-strength-of-de-identified-data/> (last visited 16/12/2020).

## **3.7 Measures to Reduce the Risk of Identification**

Based on the analysis of the factors that influence the risk of identification, the present sub-section discusses possible measures to render identification more difficult or ideally prevent it altogether. These measures are the building blocks later used in pseudonymization and anonymization.

The discussion is mostly structured along the factors that influence risk that were identified in the previous subsection. The factors without obvious measures accessible to controllers are left out, however.

### **3.7.1 Restricting Access to Personal Data**

For actors to identify natural persons behind personal data, they require access to this data. Restricting such access can thus prevent all excluded actors from attempting identification.

Note that this goes hand in hand with access control required by the principle of *purpose limitation*. Here, access to the data is only then justified when it is necessary for the declared purposes of the processing activity.

### **3.7.2 Avoiding access by any actors motivated to identify data subjects**

Access control to the personal data restricts access to only authorized parties. These can include the following:

- The controller (i.e., an organization) who may also conduct separate processing activities that, according to the principle of purpose limitation, should not gain access to the personal data at hand;
- potential processors;
- natural persons working under the authority of the controller or processor (i.e., employees);
- potential third party recipients with whom the data is shared.

Where any of these parties has an obvious motivation to identify data subjects for other purposes than those of the authorized processing, measures should be taken to avoid this. The following examples illustrate this:

- Potential processors with an obvious motivation to identify data subjects for other purposes should be avoided.
- Employees should be made aware (e.g., through training) or contractually obliged to identify data subjects only when this is necessary for the authorized purposes.

### **3.7.3 Avoiding access by any actors who may already have information about data subjects**

As was shown in Figure 5 above, identification requires linking to the representation of persons that are known, i.e., already present in the information that is already in possession of the actor. Such information can be either data or knowledge.

This means for example that it may be better to avoid a process or recipient who already is in possession of different kinds of data of the same population of data subjects.

Also, this means that where possible, it should be avoided that employees handle data of persons they may potentially know. A large organization could for example organize work in a manner where employees in one part of the country handle cases of data subjects from another part of the country. This evidently reduces the risk that an employees would handle the data of someone known to them.

### **3.7.4 Prevent the coming together of the personal data with additional information**

Indirect identification of data subjects requires that the personal data comes together with additional information to be linked together. The linking of data typically requires a computing platform.

A controller can prevent this

- by preventing that the personal data can leave the system dedicated to its processing (e.g., as a copy on a USB stick) and
- by preventing external (additional) data to be loaded on internal systems (e.g., from a USB stick or a download from the Internet) and preventing unauthorized software (that is used for linking) to be installed and executed.

### **3.7.5 Reducing the identification potential of the data itself**

This multi-faceted measure is usually the most prominent when discussing *pseudonymization* and *anonymization*. The following description is based on the model of *identification* provided by Figure 5 above (see Section 3.5), the distinction of different information elements with relevance to identification (see Section 3.2), and the analysis of linking (see Section 3.4). The aspect of identification that is relevant here is that of linking of two (information or) data sets. Different kinds of information elements in the data sets permit different kinds of linking. This is why the following discussion is structured according to kinds of information elements.

For each kind of information element, the discussion describes possible measures that impede or ideally prevent that kind of linking. It may also point out risks to the efficiency of such measures, sometimes in the form of known “attacks”.

In all the described scenarios, the linking takes place between two sets of data (or information). One of these is in possession of the actor who identifies; the other constitutes the personal data (or information) for which identification shall be impeded or ideally prevented. The presented measures modify this latter data set in ways that reduce its potential of identification. The latter data set, in its state before such modification, will be called *original data set* in the following discussion.

A detailed more detailed discussion of the topic by S. Garfinkel was published by NIST<sup>53</sup>.

#### **3.7.5.1 Deterministic linking of unique handles**

The most straightforward manner of linking records of two independent data sets is deterministic linking based on the comparison of unique handles. For this to work, both data sets obviously need to contain handles belonging to the same identity domain. The objective of preventing such linking is therefore to avoid that the data set contains any handles from identity domains used elsewhere.

---

<sup>53</sup> Simson L. Garfinkel, De-Identifying Government Datasets, NIST Special Publication 800-188 (2<sup>nd</sup> Draft), 2016, <https://csrc.nist.gov/publications/detail/sp/800-188/draft> (last visited 6/1/2021).

Starting from an original data set that may contain unique handles from other identity domains, there are three ways of eliminating this:

- (i) Deletion of all unique handles from the data set;
- (ii) Replacement of unique handles with unique handles from a newly created identity domain.

Deletion evidently prohibits linking. Also the replacement of unique handles with ones that are newly created within a new identity domain prevents any linking on equality to other data sets.

The latter option is often used when data sets are structured and unique handles are used to represent structure in the data. This is for example the case in relational data bases where the data set consists of multiple tables whose relations are based on the unique identifiers. In that case, the unique identifiers typically take the role of *primary* or *secondary keys*.

The replacement of unique handles can take two strategies. Namely, the newly created unique handle can be:

- **independent** of original unique handles;
  - For example, random numbers.
- **derived from** the original **unique handles**.
  - In a manner that **allows inversion**.
    - For example, through **encryption** of a unique handle where the inversion is the decryption of the new handle.
  - In a manner that **does not allow inversion**.
    - For example when using a one-way function with a **secret key**, such as an HMAC.

A more detailed discussion of options can be found in the ENISA report on pseudonymization<sup>54</sup>.

Evidently, the use of unique handles from an identity domain that is not used anywhere else prevents deterministic linking. The effectiveness of this prevention is not always given, however. In certain situations, it is possible to perform certain kinds of linking anyhow. Such situations are mostly due to flaws in how the new unique handles are created. The following examples shall illustrate this:

- Predictable systematics (i.e., deviation from random behaviour) in schemes to create independent new unique handles can often be exploited to permit linking. A simple example is the use of serial numbers as new unique handles. Here, additional knowledge on when the data pertaining to a person were collected can significantly aid to determine the new unique handle assigned to the person. For example, knowing that a data subject was one of the first participants means that the assigned new unique handle (i.e., serial number) must be in a certain range. A similar situation can arise when new unique handles are created by a predictable sequence of a pseudo-random number generator.

---

<sup>54</sup> See for example ENISA, Recommendations on shaping technology according to GDPR provisions; An overview on data pseudonymisation, November 2018, Contributors: Konstantinos Limniotis(Hellenic DPA), Marit Hansen(DPA Schleswig-Holstein), Editors: Athena Bourka(ENISA), Prokopios Drogkaris(ENISA), ISBN 978-92-9204-281-3, DOI10.2824/74954.

- Creation schemes for new unique handles based on an original unique handle that use a known or guessable one-way-function (i.e., a function that is not invertible, such as *sha1*) without use of a secret. This permit “attackers” to compute the new unique handle in cases where they know the original unique handle of a person. Obviously, that allows linking and identification of the person behind the according data record.
- Creation schemes for new unique handles based on an original unique handle that use a known one-way-function can under certain conditions be inverted by using brute force. This is typically the case when the set of all original unique handles is known and contains only a limited number of elements. This renders it possible to compute the pseudonym for every possible handle. It basically creates a lookup table that inverts the pseudonym creation scheme and therefore permits identification of all data subjects through deterministic linking on the original unique handle. A well-known example of such an inversion attack involved pseudonymized identifiers of taxies<sup>55</sup>.

### 3.7.5.2 Linking of quasi-identifiers

Another very common way of linking two data sets is by performing deterministic or probabilistic linking based on quasi-identifiers. Such linking is not guaranteed to be unique for all data records and probabilistic linking usually leaves a certain uncertainty, but typically, such linking can be used to link and thus identify a significant subset of data subjects.

To impede or prevent such identification, the uniqueness of the quasi-identifiers must be reduced. The most common measures used to achieve this are the following:

- **Deletion** of parts or the whole of an quasi-identifier;
- **Generalization** of the values that the quasi-identifier is composed of.

The former is evidently effective since data subject that originally distinguished themselves based on a now deleted value become undistinguishable. Or in other words, the remainder of the quasi-identifier after deletion of elements has become less unique.

The latter measure of generalization is based on the idea of reducing detail in the data and result in a “coarser” data set such that distinctions of data subjects based on details are no longer possible. More precisely, generalization maps multiple possible original values to a single “coarser” value. The objective is that multiple modified quasi-identifiers map to a single, coarser, value and thus make data subjects indistinguishable from one another. This is illustrated by the following examples:

- To generalize a ratio-scale<sup>56</sup> value, an interval of original values is mapped to a single output value. For example, sets of 356 possible dates of birth are mapped to a single year of birth. Similarly, it is possible to map the age of a person to a “generation” such as *baby boomers*, *generation X*, and *millennials*<sup>57</sup>. The latter illustrates that the intervals do not need to be regular.

---

<sup>55</sup> Vijay Pandurangan, On Taxis and Rainbows: Lessons from NYC’s improperly anonymized taxi logs, blog entry, June 22, 2014, <https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1> (last visited 24/11/2020).

<sup>56</sup> See for example [https://en.wikipedia.org/wiki/Level\\_of\\_measurement#Ratio\\_scale](https://en.wikipedia.org/wiki/Level_of_measurement#Ratio_scale) (last visited 25/11/2020).

<sup>57</sup> See for example [https://en.wikipedia.org/wiki/Generation#Western\\_world](https://en.wikipedia.org/wiki/Generation#Western_world) (last visited 25/11/2020).



- Ordinal-scale<sup>58</sup> values can be generalized by grouping adjacent values. A common example for this are 5-digit ZIP codes that are grouped depending on their first two digits. For example, the ZIP code *04609* of Bar Harbor, Maine, could be mapped to *04\*\*\**.
- Nominal-scale<sup>59</sup> values can be generalized by forming categories. For example, a person's nationality such as *Italian, Spanish, German*, etc. could be assigned to the category *European*.
- It is also possible to generalize multiple attributes together. For example, two attributes create a two dimensional space of possible values. To generalize the two-dimensional values, this space can then be partitioned into areas. This is equivalent to defining intervals in a single dimension. It is illustrated in the following Figure 8 that was taken from Kristen LeFevre et al.<sup>60</sup>

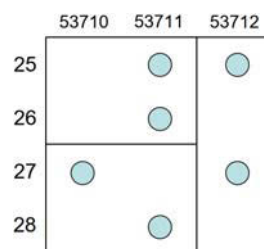


Figure 8: Example of a two-dimensional generalization with ZIP code and age by LeFevre et al.

The most common method to assess whether deletion and generalization in quasi-identifiers sufficiently impedes linkability is *k-anonymity*<sup>61</sup> by Samarati and Sweeney. In particular, the method consists of verifying that every generalized quasi-identifier occurs at least *k* times in the data set. This evidently introduces ambiguity into the possible linking. Any link attempt yields at best a set of *k* undistinguishable candidates for the matching data subject.

When for a chosen *k*, *k-anonymity* has not been reached, there are two options for how to proceed:

- Modify the generalization in a way that *k-anonymity* can be reached. This can for example be done by changing interval boundaries or categorizations.
- Delete the data records whose generalized quasi-identifiers fail to reach the *k*-threshold. This is sometimes called *record suppression*<sup>62</sup>.

Note that while generalization together with *k-anonymity* is indeed an effective measure to impede linking on quasi-identifiers, even unique linking on other data elements of the data set may still be possible. The use of the term *anonymity* in *k-anonymity* may therefore give a wrong impression. Unique linking may for example be possible based on a unique data value (an age of 117 years) or a unique combination of values that are not part of the quasi-identifier. Also, the sub-section on *attribute linking* above (see section 3.4.1) described a *homogeneity attack*, where the *k* or more data

<sup>58</sup> See for example [https://en.wikipedia.org/wiki/Level\\_of\\_measurement#Ordinal\\_scale](https://en.wikipedia.org/wiki/Level_of_measurement#Ordinal_scale) (last visited 25/11/2020).

<sup>59</sup> See for example [https://en.wikipedia.org/wiki/Level\\_of\\_measurement#Nominal\\_level](https://en.wikipedia.org/wiki/Level_of_measurement#Nominal_level) (last visited 25/11/2020).

<sup>60</sup> See Figure 4c on page 4 in Kristen LeFevre, David J. DeWitt and Raghu Ramakrishnan, Multidimensional K-Anonymity, Technical Report 1521, Department of Computer Sciences, University of Wisconsin, Madison, Revised June 22, 2005, <https://ftp.cs.wisc.edu/pub/techreports/2005/TR1521.pdf> (last visited 16/12/2020).

<sup>61</sup> See footnote 45 above.

<sup>62</sup> See for example 2<sup>nd</sup> paragraph on page 3 in Garfinkel, Simson & Abowd, John & Martindale, Christian. (2019). Understanding database reconstruction attacks on public data. Communications of the ACM. 62. 46-53. 10.1145/3287287, <https://ecommons.cornell.edu/handle/1813/89104> (last visited 22/12/2020).

records with the same generalized quasi-identifier share the same attribute value and therefore, the linking of that attribute value to the data subjects becomes possible.

### 3.7.5.3 Linking of identity-relevant properties

In addition to linking based on unique handles and quasi-identifiers, linking is also possible based on unique combinations of values of identity-relevant properties. More precisely, the following section considers both, identity-relevant properties together with (the possibly already generalized) quasi-identifiers.

Also here, the idea behind impeding or preventing linking is based in reducing the uniqueness of data records.

The following discussion is based on a literature review of technical methods to achieve this. Keywords for this literature include among others *anonymization*, *de-identification (and re-identification)*, *disclosure control*, and *privacy preserving publishing*. It is impossible here to provide a comprehensive overview of the wealth of technical methods found in the literature; there are too many methods and many of them come in different variations and combinations. For this reason, the following attempts to provide a categorization of the abstract concepts of transformation that underlie these technical methods. These concepts of transformation describe a way of modifying the original data in order to impede linking. Actual methods are then either implementations of a single concepts or a combination of several concepts.

When looking at data as a model of the world, these concepts of transformation have a certain effect on these models. At the highest level of the categorization, this view permits to distinguish two kinds of concepts:

- Concepts that result in “truthful”, yet less detailed, models of the world, and
- concepts that result in models of the world that deviate from the truth but are close to the truth and possibly even share certain properties with the truth.

This distinction is used to structure the discussion of concepts of transformation. For a good alternative overview of “anonymization operations”, see Fung et al<sup>63</sup>.

#### 3.7.5.3.1 Truthful concepts of transformation

The following describes truthful concepts of transformation.

##### 3.7.5.3.1.1 Deletion

This concept of transformation is also called *suppression* and *non-disclosure*. It reduces detail in the original model by leaving away certain information; the remaining data constitute a truthful model.

Deletion can affect different data elements:

- **A single attribute** belonging to a **single data subject**:  
This may for example be done, when a value of an attribute becomes too rare. In this case, the value is often replaced by an asterisk “\*”. An example is the age of a person. A very

---

<sup>63</sup> Fung, Benjamin & Wang, Ke & Fu, Ada & Yu, Philip, 2010, Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques, DOI 10.1201/9781420091502, [https://www.academia.edu/24652325/Introduction\\_to\\_Privacy\\_Preserving\\_Data\\_Publishing](https://www.academia.edu/24652325/Introduction_to_Privacy_Preserving_Data_Publishing) (last visited 23/12/2020).

high age is quite rare and therefore identifying. In 2020, the oldest person alive dies at the age of 112 years old<sup>64</sup>. The oldest known person ever died at the age of 112 and according to Wikipedia<sup>65</sup>, there are only 50 persons in history who became 115 or older. It may therefore be a good idea to suppress age values that lie above a certain threshold. (Note as an alternative to deletion, *top-coding* is also used. Since it deviates from the truth, it is described below).

- **A single attribute across all data subjects:**  
It is also possible to delete a single data element for all data subjects in the data set. In tabular data, this would delete a whole **column**. This is typically done when this data element is considered to be highly identifying. Location data that is known to be almost impossible to anonymize may be a candidate for such deletion.
- **All attributes belonging to a single data subject:**  
It is also possible to delete individual data records. This may for example be used when certain data subjects are easily recognized due to a very rare and identifying combination of values.
- **All attributes belonging to a group of data subjects:**  
It is possible to delete all data records belonging to a group of data subjects. This is often called *cell suppression*. A typical use case comes from the application of k-anonymity where quasi-identifiers are generalized such that each generalized quasi-identifier value occurs at least k times. The data records sharing the same quasi-identifier value are often called *cells*. If the chosen generalization method works for most of the data set, but leaves one or few cells of a size smaller than k, then cell suppression may be used to reach k-anonymity for the data set anyhow.
- **Resampling** of time series:  
A special case of deletion can be used in time series of a given attribute. It only keeps selected values, while deleting the others.

Note that if a suppression is made explicit in the data (for example through the use of “\*”) or an actor has additional knowledge to detect suppression, this can by itself reveal something about the data subject. Assume for example, that an attribute of *gender* is suppressed where its value is not *male* or *female*. Evidently, in this scenario, an “\*” can convey highly sensitive information about a data subject.

### 3.7.5.3.1.2 Generalization

Generalization was already discussed above for quasi-identifiers. The same transformation concept can be applied also to identity-specific properties. The following discussion of the concept is somewhat more systematic and comprehensive than its treatment for quasi-identifiers. The section also provides additional examples that go beyond those of generalization of quasi-identifiers.

Like in the case of quasi-identifiers, the concept of generalization maps multiple original values to a single generalized value. Data subjects that can be distinguished based on different original values then become indistinguishable at the generalized level where the original values map to the same generalized value.

---

<sup>64</sup> <https://www.guinnessworldrecords.com/news/2020/5/worlds-oldest-man-bob-weighton-dies-aged-112> (last visited 18/12/2020).

<sup>65</sup> [https://en.wikipedia.org/wiki/List\\_of\\_the\\_verified\\_oldest\\_people](https://en.wikipedia.org/wiki/List_of_the_verified_oldest_people) (last visited 18/12/2020).

The following distinguishes different kinds of generalization:

(i) Generalizations that apply **coarser scales of measurement to one or several attributes of single data subjects**:

One way of generalizing data is to change the attributes to a coarser scale of measurement. This mapping loops over all data subjects, taking one or several attributes of a single data subject as input at each round.

In the most simple case, where a scale of measurement is defined by regular intervals of values, this can be seen as a change in the units of measurement. For example, instead of measuring a distance at millimeter precision, it can now be measured only at centimeter precision. This simple concept can be applied more generally also to more complex cases. This is illustrated in the following examples:

- In the case of regular intervals, generalization can be seen as a rounding of the original value to a lower precision value. For example, a precise person height (172,54 cm) can be rounded to the centimeter (173 cm). Evidently, many different detailed values then map to the same rounded value.
- Nominal-scaled values can be generalized by forming categories. For example, the International Classification of Diseases<sup>66</sup> uses a categorization of medical symptoms and signs to categories such as “*symptoms involving cardiovascular system*”, “*symptoms involving respiratory system*” and *other chest symptoms*”, etc.<sup>67</sup>
- A location originally described by a coordinate pair (e.g., latitude and longitude) can be generalized to a position described by ZIP code area, census district, province, country, or continent. Note that this maps a multitude of different pairs of ratio-valued attributes into a single nominal one.
- The generalization of a list of languages that a person can speak fluently could map to the two categories of *monolingual* and *multi-lingual*.

These kinds of generalization can thus be defined by a mapping from the original scale of measurement to a coarser scale of measurement. More generally, it consists of a partition<sup>68</sup> of the set of all possible original values of the input attribute(s) into a multitude of subsets. All original values in a given subset are then mapped to the same coarser value.

In the same way as with the generalization of quasi-identifiers, the generalization of identity-specific properties does not guarantee that linking can be prevented. It is evident that this depends on how many data subjects end up with a given coarser attribute value. If only a single data subject ends up with a given coarser attribute value, linking is still possible and the data subject can therefore still be identified. Even if every no coarser attribute value is shared by less than k data subjects, combinations of attribute values can still be unique for a single data subject.

---

<sup>66</sup> See for example [https://en.wikipedia.org/wiki/International\\_Classification\\_of\\_Diseases](https://en.wikipedia.org/wiki/International_Classification_of_Diseases) (last visited 25/11/2020).

<sup>67</sup> See for example [https://en.wikipedia.org/wiki/List\\_of\\_ICD-9\\_codes\\_780%E2%80%9393799:\\_symptoms,\\_signs,\\_and\\_ill-defined\\_conditions](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_780%E2%80%9393799:_symptoms,_signs,_and_ill-defined_conditions) (last visited 25/11/2020).

<sup>68</sup> A partition is a subdivision of a set into subsets such that these subsets do not overlap and that the union of all subsets is equal to the original set.

(ii) Generalizations that maps **multiple attributes of a single data subject** to coarser **statistical attributes**:

This form of generalization again loops over every single data subject and takes multiple attributes of this person as input to yield a coarser statistical description of these values. The following example shall illustrate this:

- A time series of a patient's body temperature could be generalized by mapping the temperature values over a single day to statistics such as the average, minimum, and maximum temperature.

This kind of generalization is defined by the applied statistics.

While statistics certainly impede the potential of linking, the same limitation apply that were discussed with the previous type of generalization.

(iii) Generalization that maps attributes of **multiple data subject** to a **single attribute describing groups of data subjects**:

This form of generalization first forms groups of data subject (typically based on the generalization of quasi-identifiers) and then loops over every group to take statistics over the distribution of attribute values in the group. This is for example applied to census data. For example, groups could be formed based on census district and gender, and each resulting group may be described by statistics such as the number of persons (count), the average age, the minimal income, etc.

It may be common to think that it is impossible to link statistical data to individual-level data sets; In other words, statistical data is free of risk of identification. As is described well by Garfinkel et al<sup>69</sup>, this is not always the case. In particular, if a multitude of statistics is available, a so called *reconstruction attack* may be possible. In their paper, Garfinkel et al provide a practical example for this. They show how in certain cases, it is possible to reconstruct original value of some or even all data subjects.

The idea behind the reconstruction attack is that the attribute values of individuals are unknowns and that each available statistical value allows to formulate an equation about these unknowns. If multiple statistical values are available, an equation system can be composed. If it is determined, it can be solved for the unknowns. Garfinkel et al show that when the equation system is underdetermined, there are multiple solutions of the unknowns that satisfy the equations. Solving systems may then be able to enumerate all possible solutions. For individual data subjects, these may all share the same value for one or more attribute. In other words, it may still be possible to find a certain solution for selected data subjects and selected attributes.

Garfinkel et al describe that in practical cases, the solution of such equation systems requires significant computing power. They argue that with the advances in computing hardware and solver methodology, reconstruction attacks have indeed become a realistic risk. They describe how this risk has determined the choice of de-identification techniques used for the U.S. 2020 census.

### 3.7.5.3.1.3 Slicing

It has been argued in section 3.2.1 that multi-dimensional data has a very high risk of containing unique combinations of attributes for data subjects. Multi-dimensional data sets therefore have a high potential for linking. *Slicing* addresses the risk inherent in multi-dimensionality.

---

<sup>69</sup> See footnote 62.

The concept of slicing takes a multi-dimensional original data set and splits it into multiple pieces, each of which being only of a small dimension. These pieces still contain individual-level data.

The linkability of records across pieces must therefore be controlled carefully. This is typically done by forming groups of data subjects (typically based on generalization of quasi-identifiers) and adding a group number as additional attribute in every piece. Typically, that results in pieces where every group contains at least a certain number (k) of data subjects. This makes the method very similar to k-anonymity since when linking a record of one piece, there are at least k matching records in each of the other pieces.

It is evident that linking on an implicit order has to be avoided too and that the order of the records in the pieces cannot remain that of the original data set. Some random shuffling to change the order may be necessary here.

Another kind of likability across pieces that needs to be avoided is to keep highly correlated data in the same piece. If they are in separate pieces, it may be possible to link record of these pieces based on this correlation. For example, assume that the original data set contains both, profession and income. Then, it may be highly likely that the only *CIO* occurring in one piece can be linked to the highest income reported in the other piece; or that that only *unemployed* person links to the lowest (or zero) income.

The concept of slicing has been proposed in multiple variations and often mixed with other concepts of transformation.

Possibly the first application of slicing in the literature is the method of **anatomization** proposed by Xiao and Tao<sup>70</sup>. It splits an original data set consisting of quasi-identifiers (e.g., age, sex, zip code) and a single attribute (e.g., disease) into two pieces that are linkable via a group number. In contrast to generalization (for k-anonymity), it uses the grouping defined by the generalization while still reporting the original values of the quasi-identifiers in one piece. Xiao and Tao compare their anatomization with generalization and show how anatomized data sets contain more useful information while providing the same protections against linking. To further, illustrate anatomization, the following figures are copied from Xiao and Tao’s paper.

tuple ID	Age	Sex	Zipcode	Disease
1 (Bob)	23	M	11000	pneumonia
2	27	M	13000	dyspepsia
3	35	M	59000	dyspepsia
4	59	M	12000	pneumonia
5	61	F	54000	flu
6	65	F	25000	gastritis
7 (Alice)	65	F	25000	flu
8	70	F	30000	bronchitis

Figure 9: Original data set (by Xiao and Tao).

---

<sup>70</sup> Xiaokui Xiao, Yufei Tao, Anatomy: Simple and Effective Privacy Preservation, VLDB 2006: 139-150, <http://www.vldb.org/conf/2006/p139-xiao.pdf> (last visited 22/12/2020).

tuple ID	Age	Sex	Zipcode	Disease
1	[21, 60]	M	[10001, 60000]	pneumonia
2	[21, 60]	M	[10001, 60000]	dyspepsia
3	[21, 60]	M	[10001, 60000]	dyspepsia
4	[21, 60]	M	[10001, 60000]	pneumonia
5	[61, 70]	F	[10001, 60000]	flu
6	[61, 70]	F	[10001, 60000]	gastritis
7	[61, 70]	F	[10001, 60000]	flu
8	[61, 70]	F	[10001, 60000]	bronchitis

Figure 10: Generalized data set (by Xiao and Tao).

row #	Age	Sex	Zipcode	Group-ID
1	23	M	11000	1
2	27	M	13000	1
3	35	M	59000	1
4	59	M	12000	1
5	61	F	54000	2
6	65	F	25000	2
7	65	F	25000	2
8	70	F	30000	2

Group-ID	Disease	Count
1	dyspepsia	2
1	pneumonia	2
2	bronchitis	1
2	flu	2
2	gastritis	1

Figure 11: Original data set sliced in two through anatomization (by Xiao and Tao).

There are various variations of the concept at hand. Susan and Christopher describe the application of the concept to higher dimensional data sets<sup>71</sup>. Onashoga et al describe the KC-slice method<sup>72</sup> that was further refined by Raju et al<sup>73</sup>.

Note that slicing does not by itself guarantee to prevent linking of data sets. But by breaking up high-dimensional data sets into multiple smaller-dimensional ones, it reduces the risk of highly identifying unique combinations.

### 3.7.5.3.2 Concepts of transformation that introduce deviations from the truth

The previous section has described transformations that are truthful, i.e., they present the data at a reduced level of detail without introducing any errors or deviations from the truth. This section provides an overview of transformations that introduce deviations from the truth.

<sup>71</sup> V.S. Susan and T. Christopher, Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes, Springerplus, 2016;5(1):964, Published 2016 Jul 4, doi:10.1186/s40064-016-2490-0, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4932023/> (last visited 23/12/2020).

<sup>72</sup> Onashoga, S. A. et al. "KC-Slice: A dynamic privacy-preserving data publishing technique for multisensitive attributes." Information Security Journal: A Global Perspective 26 (2017): 121 – 135,

<sup>73</sup> N.V.S. Laskshmipathi Raju & M.N. Seetaramanath & Rao, P. Srinivasa Rao. (2018). An enhanced dynamic KC-Slice model for privacy preserving data publishing with multiple sensitive attributes by inducing sensitivity. Journal of King Saud University - Computer and Information Sciences. 10.1016/j.jksuci.2018.09.013, <https://www.sciencedirect.com/science/article/pii/S1319157818304324> (last visited 23/12/2020).

### 3.7.5.3.2.1 Top- and bottom-coding

As described above, one way of identifying persons is based on rare and thus unique values. Typical examples are a very old age and a very high income. Similarly, a very short body height may be rare or unique (in a given dataset).

Top- and bottom-coding avoids identification based on rare very high or very low ratio values, respectively. For this purpose, a threshold is chosen and every value higher or lower than the threshold, respectively, is replaced by the threshold value. For example, 90 may be chosen as a threshold age and all age values greater than 90 in the data set are replaced by 90.

Evidently, this introduces errors in the model of the world. For example, it introduces an error in the calculation of the average age. Top- and bottom-coding are routinely used in statistical publications such as census data<sup>74</sup>.

### 3.7.5.3.2.2 Data swapping

Another non-truthful transformation is *data swapping*. The basic concept is that data values are randomly swapped between individuals contained in a data set. Typically, such swapping is restricted to individuals belonging to the same subsets of data. The subsets are typically those used in the generalization of quasi-identifiers and are often called "group" or "cell".

Data swapping impedes linking by changing the combinations of values that could be unique and identify a person. While this introduces deviations from the truth, the method aims at keeping certain characteristics (typically of a cell, i.e., a group of persons) invariant. Such characteristics include the distribution of values or the average, median, etc.

For a more detailed discussion of data swapping, see for example Fienberg and McIntyre<sup>75</sup>. A description how data swapping was used in the U.S. 1990 census is provided by McKenna<sup>76</sup>.

---

### 3.7.5.3.2.3 Random noise injection

An important non-truthful transformation is *noise injection* (aka. *noise addition*). Here, a random error is added to truthful data. The more error is added, the less likely that identification is still possible. This is most evident by the probabilistic linking of continuous property values based on similarity. The more noise is injected, the lower the similarity of the values.

The key question with noise injection is how much noise needs to be added to prevent identification. The probably best-known approach to answering this question is *differential privacy* that was first

---

<sup>74</sup> See for example footnote 53 on page 3.

<sup>75</sup> Fienberg, S. and J. McIntyre. "Data Swapping: Variations on a Theme by Dalenius and Reiss." *Privacy in Statistical Databases* (2004), <http://www.stat.cmu.edu/~fienberg/DLPapers/Fienberg-McIntyre-LNCS-2004.pdf> (last visited 11/1/2021).

<sup>76</sup> Laura McKenna, 2018. "Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing," Working Papers 18-47, Center for Economic Studies, U.S. Census Bureau. <https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20Techniques%20for%20the%201970-2010%20Censuses.pdf> (last visited 11/1/2021).



proposes by Dwork et al.<sup>77</sup>. According to Sartor<sup>78</sup>, “[m]any privacy researchers regard [differential privacy] as the ‘gold standard’ of anonymization”. This may largely be since “it offers a guaranteed bound on loss of privacy due to release of query results, even under worst-case assumptions”<sup>79</sup>. Referencing Dwork et al.<sup>80</sup>, Wikipedia states: “Although it does not directly refer to identification and re-identification attacks, differentially private algorithms probably resist such attacks.” This statement seems to be further supported by McClure and Reiter<sup>81</sup>.

Sartor provides a good introduction to the topic; a more detailed introduction was provided by Wood et al.<sup>82</sup>. Sartor lists further introductory resources to the topic.

Differential privacy is not a single transformation to reduce the identification potential of a data set. Much rather, differential privacy is a mathematical framework that is based on a mathematical definition of what privacy actually is. According to Sartor, “[t]here are now hundreds of published differentially private mechanisms” for which there are mathematical proofs that they comply with the mathematical framework. He provides the examples of building a histogram<sup>83</sup>, taking an average<sup>84</sup>, releasing micro-data<sup>85</sup> (i.e., individual-level data), and generating a machine learning model<sup>86</sup>.

The definition of privacy given by *differential privacy* is based on the idea that it should not be possible to determine whether a given individual is contained in a data set. This is done by comparing the data set that contains a given individual with one that does not. Ideally, if there is no difference at all, it is obviously not possible to determine whether the person is reflected in the data set. This is not possible, however. Therefore, differential privacy requires that the difference must be very small. As small numbers are often represented by an  $\epsilon$ , it is also called  $\epsilon$ -*differential privacy*.

---

<sup>77</sup> Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2017. “Calibrating Noise to Sensitivity in Private Data Analysis”. *Journal of Privacy and Confidentiality* 7 (3):17-51. <https://doi.org/10.29012/jpc.v7i3.405>.

<sup>78</sup> Nicolas Sartor, Explaining Differential Privacy in 3 Levels of Difficulty, aircloak blog, <https://aircloak.com/explaining-differential-privacy/> (last visited 13/1/2021).

<sup>79</sup> Hsu, Justin & Gaboardi, Marco & Haeberlen, Andreas & Khanna, Sanjeev & Narayan, Arjun & Pierce, Benjamin & Roth, Aaron. (2014). Differential Privacy: An Economic Method for Choosing Epsilon. Proceedings of the Computer Security Foundations Workshop. 2014. 10.1109/CSF.2014.35. <https://arxiv.org/abs/1402.3329> (last visited 15/1/2021).

<sup>80</sup> See footnote 77.

<sup>81</sup> McClure, D. and J. Reiter. “Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data.” *Trans. Data Priv.* 5 (2012): 535-552, <http://www.tdp.cat/issues11/tdp.a093a11.pdf> (last visited 15/1/2021).

<sup>82</sup> Wood, Alexandra, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, et al. 2018. Differential Privacy: A Primer for a Non-Technical Audience. *Vanderbilt Journal of Entertainment & Technology Law* 21 (1): 209. <http://nrs.harvard.edu/urn-3:HUL.InstRepos:38323292> (last visited 13/1/2021).

<sup>83</sup> Dwork C. (2008) Differential Privacy: A Survey of Results. In: Agrawal M., Du D., Duan Z., Li A. (eds) *Theory and Applications of Models of Computation*. TAMC 2008. Lecture Notes in Computer Science, vol 4978. Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-540-79228-4\\_1](https://doi.org/10.1007/978-3-540-79228-4_1), [https://web.cs.ucdavis.edu/~franklin/ecs289/2010/dwork\\_2008.pdf](https://web.cs.ucdavis.edu/~franklin/ecs289/2010/dwork_2008.pdf) (last visited 13/1/2021).

<sup>84</sup> Nozari, Erfan, P. Tallapragada and J. Cortés. “Differentially Private Average Consensus with Optimal Noise Selection.” *IFAC-PapersOnLine* 48 (2015): 203-208. <http://www.ee.iisc.ac.in/people/faculty/pavant/files/papers/C10.pdf> (last visited 13/1/2021).

<sup>85</sup> Raffael Bild, Klaus A. Kuhn, Fabian Prasser, SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees, *Proceedings on Privacy Enhancing Technologies*, 2018(1), 67-87, <https://doi.org/10.1515/popets-2018-0004> (last visited 13/1/2021).

<sup>86</sup> Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, Li Zhang, Deep Learning with Differential Privacy, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (ACM CCS)*, pp. 308-318, 2016, DOI 10.1145/2976749.2978318, <https://arxiv.org/abs/1607.00133v2> (last visited 13/1/2021).

According to Sartor, “[i]n real life scenarios, the inventors of the concept suggested to keep the epsilon between 0.1 and 1.” More detailed information about choosing an adequate  $\epsilon$  is given by for example by Hsu et al<sup>87</sup>.

Differential privacy cannot only quantify the privacy loss of a single disclosure of a data set, but is also capable of modelling the overall privacy loss in a composition of multiple disclosures. This is for example possible in the case of multiple queries to a data base or the publication of a multitude of statistics by a census bureau. Each individual disclosure increased the privacy loss. Differential privacy makes it therefore possible to define a “privacy budget” that gets depleted with every disclosure. When this budget is spent, further disclosures must then be avoided.

Differential privacy is a complex topic and requires a high level of mathematical skill to be understood and thus used. The Linknovate Team has conducted a survey in 2018 and found that only very large players are typically engaged in differential privacy activities<sup>88</sup>. Among the most prominent practical applications of differential privacy is the Census 2020<sup>89</sup> by the U.S. Bureau of the Census and the mining of user data by Apple<sup>90 91</sup>. But neither of these has yet proven to be success stories.

The Task Force on Differential Privacy for Census Data of the University of Minnesota’s Institute for Social Research and Data Innovation has voiced doubts whether differential privacy was a good choice, suggesting that the method was overkill and may compromise the utility of the census data. They state that they “believe that the differential privacy approach is inconsistent with the statutory obligations, history, and core mission of the Census Bureau.”

Also the application by Apple was met with critique. In particular, Tang et al found that the  $\epsilon$  used in was way too high to guarantee privacy (up to 16 instead of a recommended maximum of 1) and that the privacy budget was renewed every day<sup>92</sup>.

To investigate the practical applicability of differential privacy, Sartor<sup>93</sup> has used a software tool for differential privacy called *PSI*<sup>94</sup> that is part of the *Harvard Privacy Tools Project*<sup>95</sup>. He writes: “Setting epsilon to 0.5, we could build user-count histograms of 3 columns and take the mean of two more. The tool estimated the 95% error on the mean to be 3%, and on the counts to be  $\pm 60$  (around 5% to 10% for most of the histogram bars). These are reasonable and useful answers, but those 5 queries

---

<sup>87</sup> Hsu, Justin & Gaboardi, Marco & Haeberlen, Andreas & Khanna, Sanjeev & Narayan, Arjun & Pierce, Benjamin & Roth, Aaron. (2014). Differential Privacy: An Economic Method for Choosing Epsilon. Proceedings of the Computer Security Foundations Workshop. 2014. 10.1109/CSF.2014.35.  
<https://arxiv.org/abs/1402.3329v1> (last visited 13/1/2021).

<sup>88</sup> Linknovate Team, Differential Privacy Leaders you Must Know, September 13, 2018,  
<https://blog.linknovate.com/differential-privacy-leaders-must-know/> (last visited 15/1/2021).

<sup>89</sup> John M. Abowd, Protecting the Confidentiality of America’s Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau, August 17, 2018,  
[https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting\\_the\\_conf.html](https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_conf.html) (last visited 15/1/2021).

<sup>90</sup> WWDC 2016. June, 2016. Engineering Privacy for Your Users.  
<https://developer.apple.com/videos/play/wwdc2016/709/>. (last visited 15/1/2021).

<sup>91</sup> WWDC 2016. June, 2016. WWDC 2016 Keynote. <https://www.apple.com/apple-events/june-2016/>. (last visited 15/1/2021)

<sup>92</sup> Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang, Privacy Loss in Apple’s Implementation of Differential Privacy on MacOS 10.12, 11 Sep 2017, <https://arxiv.org/abs/1709.02753> (last visited 15/1/2021).

<sup>93</sup> See footnote 78.

<sup>94</sup> <http://psiprivacy.org/static/about/> (last visited 15/1/2021).

<sup>95</sup> <https://privacytools.seas.harvard.edu/> (last visited 15/1/2021).

exhausted the budget. Were one to adhere strictly to differential privacy, that database (a small portion of the California Demographic Dataset) could never be queried again.”

In addition, Sartor used the *ARX Data Anonymization Tool*<sup>96</sup> to de-identify a table with individual-level data that contains 16 attributes for 5369 persons, choosing a high  $\epsilon$  value (of 2, i.e., double the maximally recommended value). The result was a table with every value replaced by a ‘\*’ symbol. While it can probably be expected that the de-identification of high-dimensional data is indeed difficult, the result is anyhow sobering.

Sartor concludes his experience by saying that differential privacy was “beautiful in theory” but a “fallacy in practice. In more detail, he writes:

“Differential privacy is a beautiful theory. If it could be made to provide adequate utility while maintaining small epsilon, corresponding complete proofs, and reasonable assumptions, it would certainly be a privacy breakthrough. So far, however, this has rarely, and arguably never happened.”

#### **3.7.5.3.2.4 Synthetic data generation**

While somewhat out of scope, the following briefly mentions synthetic data generation. It can be questioned whether this is indeed a transformation of original data that results in a data set with a lesser identification potential. The basic idea is to synthetically (randomly) generate data that share certain statistical properties with the original data. Such properties could be averages or medians, as well as distributions of values within the dataset. These properties are then extracted from the original data set and the only truthful information that is disclosed. From the point of view of the potential of identification, such synthetic data are equivalent to a data set that just publishes the preserved statistical properties.

#### **3.7.5.4 Summary of transformations that reduce the identification potential of data**

This section provides a summary of the above presented transformation concepts to reduce the identification potential of data sets. It points out some important characteristics that help understand how to apply the transformations and what guarantees they can provide that identification is no longer possible.

Most of the above described concepts of transformation fail to consider the data set as a whole but rather has a **limited scope**. For example, top-coding considers only a single attribute value of a single individual, and generalization in the context of k-anonymity focuses exclusively on the linking of the data elements that compose a quasi-identifier (while leaving all the other data elements unaffected).

Transformations of data sets lead to a **gradual reduction of their identification potential**. In particular, generalization gradually reduces the level of detail contained in the data and noise injection gradually increases the level of error added to the data. Accordingly, the majority of transformations is parameterized (for example with k in k-anonymity,  $\epsilon$  in  $\epsilon$ -differential privacy, or the threshold chosen for top-coding) and in practice, typically a multitude of transformations is applied to the data set. Consequently, depending on the choice of the combination of transformation and their parameters, it is possible to reduce the identification potential of the data set almost gradually in small steps. This holds for both, direct and indirect identification.

---

<sup>96</sup> <https://arx.deidentifier.org/> (last visited 15/1/2021).

The key question in this situation is **how much** along gradual scale a data set has to be transformed in order **to prevent direct** and **indirect identification**, respectively. Unfortunately, there is a lack of methods that could yield any certain answers to this question. While there are some “privacy models” that attempt to address the question, they all are limited in scope and focus on just one of many ways of linking. For example, k-anonymity “measures” whether linking on quasi-identifiers can be used to identify a person. It fails to address any other kind of linking that could lead to an identification, as for example the linking on identity-relevant properties.

As is evident in the presented model of identification (see Figure 5), whether direct and indirect identification is possible always depends on the information that is directly available to the actor or that can be obtained indirectly from external sources. What information has to be considered can be difficult to determine. For example, it may be difficult for controllers to establish whether employees who work with the personal data actually know data subjects and are therefore able to identify them by recognizing combinations of attribute values. Similarly, it is very difficult to establish what external data sets may exist that could be used for indirect identification. This is further aggravated by the fact that when determining identifiability, also future developments have to be taken into account<sup>97</sup>.

The most promising method to determine whether a data set permits identification of data subjects is probably  $\epsilon$ -differential privacy. While it fails to address identification directly, it refrains from making any assumptions on what (additional) information is available to “attackers”. It thus is likely the only method that comes close to providing guarantees that identification is not possible. The limitation is then still the question of what  $\epsilon$  identification is no longer possible. Another limitation is if multiple actors publish differentially private data about the same attributes without coordinating a common “privacy budget”.

In practice, there seem to be two common ways to determining whether a data set still permits identification (and can thus be disclosed/published)<sup>98</sup>:

- Rule-based and
- principles-based

(statistical) disclosure control.

In the former case of rule-based disclosure control, “a rigid set of rules is used to determine whether or not the results of data analysis can be released.”<sup>99</sup> Examples for such rules include the following<sup>100</sup>:

- “A table may only be released if there are at least three observations for each cell”,
- “A regression may be released if not based entirely on categorical data”,
- “A Herfindahl index of over 0.3 should only be released as ‘over 0.3’”, and
- “Variance-covariance matrices  $X'X$  may not be released”.

In the latter case of principles-based disclosure control, the decision whether the data set is “safe to be disclosed” is based on the subjective assessment of risk by the researcher(s) and “output

---

<sup>97</sup> See Recital 26 sentence 4 GDPR.

<sup>98</sup> See for example, Ritchie, Felix & Elliot, Mark. (2014). Principles- Versus Rules-Based Output Statistical Disclosure Control In Remote Access Environments. IASSIST Quarterly, 39, DOI 10.29173/iq778, [https://www.researchgate.net/publication/273725328\\_Principles- Versus Rules- Based Output Statistical Disclosure Control In Remote Access Environments](https://www.researchgate.net/publication/273725328_Principles- Versus Rules- Based Output Statistical Disclosure Control In Remote Access Environments) (last visited 21/1/2021).

<sup>99</sup> Wording taken from Wikipedia, [https://en.wikipedia.org/wiki/Statistical\\_disclosure\\_control#Rules- Based\\_SDC](https://en.wikipedia.org/wiki/Statistical_disclosure_control#Rules- Based_SDC) (last visited 21/1/2021).

<sup>100</sup> Copied from Ritchie et al, see footnote 98.

checker(s)", both assumed to be trained in disclosure control. While the decision may be supported with "rules of thumb", these rules never prevent a disclosure and the responsibility lies uniquely on the shoulders of the researcher(s) and output checker(s).

This situation often prevents certainty in the answer to whether (direct or indirect) identification of a data set is still possible. There will always be a level of uncertainty. This situation becomes even more complex when considering future developments. A data set that is "safe to disclose" at the point of time of publication, may become identifiable at a later point in time. Hornung and Wagner describe how identifiability can occur suddenly or gradually<sup>101</sup>. They reason that in particular in the latter case, that the situation is not addressed in the GDPR and that this causes a legal uncertainty that may yet have to be closed by the legislator.

### 3.7.5.5 Tools for reducing the identification potential of personal data

Implementing and applying transformations to reduce the identification potential of data can be difficult and time consuming. Most practitioners will therefore likely use already available software tools. The following provides some starting points for the search of suitable tools.

Overviews of existing tools have been provided by a multitude of players (see links to overviews in the footnotes):

- U.S. National Institute of Standards and Technology (NIST)<sup>102</sup>,
- Aircloak<sup>103</sup>,
- Johns Hopkins University<sup>104</sup>,
- YourTechDiet<sup>105</sup>, and
- Electronic Health Information Laboratory (EHIL)<sup>106</sup>,

Note that some of the available tools must rather be considered to be tool boxes since they implement a variety of transformation concepts and algorithms. For example, the open source *ARX Data Anonymization Tool* by the Technical University of Munich contains both, tools and a programming library that support a multitude of privacy models including k-Anonymity,  $\ell$ -Diversity, t-Closeness, and differential privacy<sup>107</sup>.

---

<sup>101</sup> Gerrit Hornung and Bernd Wagner, Der schleichende Personenbezug. Die Zwickmühle der Re-Identifizierbarkeit in Zeiten von Big Data und Ubiquitous Computing, Computer und Recht 2019, 565-574, <https://www.cr-online.de/60002.htm>, (in German), (last visited 21/1/2021).

<sup>102</sup> <https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/focus-areas/de-id/tools> (last visited 21/1/2021).

<sup>103</sup> <https://aircloak.com/top-5-free-data-anonymization-tools/> (last visited 21/1/2021).

<sup>104</sup> <https://dataservices.library.jhu.edu/resources/applications-to-assist-in-de-identification-of-human-subjects-research-data/> (last visited 21/1/2021).

<sup>105</sup> <https://www.yourtechdiet.com/blogs/6-best-data-anonymization-tools/> (last visited 21/1/2021).

<sup>106</sup> <http://www.ehealthinformation.ca/fag/de-identification-software-tools/> (last visited 21/1/2021).

<sup>107</sup> <https://arx.deidentifier.org/overview/privacy-criteria/>

## 4 Pseudonymization

Pseudonymity is concerned with the possibility of direct identification of data subjects in a data set. This contrasts with anonymity that is in addition concerned also with indirect identification. Pseudonymous data is personal data where direct identification of data subjects is rendered impossible.

This section gives an overview of *pseudonymization* by covering the following arguments:

1. An introduction.
2. An analysis of the definition of *pseudonymization* given in the GDPR.
3. A description of the context in which pseudonymization is embedded.
4. Some typical usage scenarios for pseudonymization.
5. The definition of important concepts related to pseudonymization.
6. The detailed description of *data pseudonymization* (which is one of these concepts).
7. A description of technical and organizational measures to be used for pseudonymization.
8. An analysis of different types of re-identification.
9. A detailed discussion of what Art. 11 GDPR states about pseudonymization.

### 4.1 Introduction to pseudonymization

The following sub-sections provide a general setting for the discussion of pseudonymization. In particular, it discusses:

- The motivation why controllers should implement *pseudonymization*,
- how *pseudonymization* reduces the risks for data subjects, and
- the importance that the GDPR assigns to the concept of *pseudonymization*.

#### 4.1.1 Motivation to use pseudonymization

This section addresses the question of why one would consider to use pseudonymization. There are three main reasons:

- (i) It is required by the GDPR,
- (ii) it reduces the risk for data subjects, and
- (iii) it permits controllers to reduce the effort of implementing technical and organizational measures.

These reasons are discussed in more detail in the following.

(i) Among the principles of data protection stated in Art. 5(1) GDPR, there are *data minimization* (Art. 5(1)(c)) and *storage limitation* (Art. 5(1)(e)). The former states that information elements that permit direct identification of data subjects can only be used when “necessary in relation to the purposes for which they are processed”. In other words, as soon as the purposes of processing do not require direct identification of data subjects, the corresponding data elements shall be deleted. This is in practice achieved by data pseudonymization. Storage limitation similarly states that “data shall be kept in a form which permits identification of data subjects for no longer than is necessary for the

purposes". When distinguishing between direct and indirect identification, this means that as soon as possible, the data shall be brought to a form that no longer permits direct identification. This is again achieved by data pseudonymization.

(ii) Recital 28 GDPR states: "The application of pseudonymisation to personal data can reduce the risks to the data subjects concerned [...]" Pseudonymization is thus a measure of risk reduction. Recital 28 continues with "[...] and help controllers and processors to meet their data-protection obligations.". This leads over to the next reason for using pseudonymization.

(iii) The GDPR takes a risk-based approach to data protection<sup>108</sup>. Accordingly, the technical and organizational measures that controllers are obliged to implement have to be proportional to the risk<sup>109</sup>. In other words, by reducing the risk through pseudonymization, a lower level of protection is required. Consequently, the effort of implementing technical and organizational measures by the controller can be reduced. This is also expressed in Recital 28 GDPR which states: "The application of pseudonymisation to personal data can [...] help controllers and processors to meet their data-protection obligations."

The reduction of risk through pseudonymization is discussed in section 4.4 below. It reasons that the pseudonymous data that is disclosed to personnel during its processing presents typically only a marginal risk. Further, *additional information* that is more critical since it permits re-identification is not processed during pseudonymization (see section 4.2) beyond being stored for further use. This renders the protection of this *additional information* relatively easy<sup>110</sup>.

The critical kind of processing that needs adequate protection is the possible re-identification. This exists from pseudonymization and newly enters into processing of identified data. The possibility to concentrate protective effort on this limited sensitive part of the overall processing allows controllers to reduce their overall effort in comparison to having to protect everything at that same level.

To illustrate the reduced effort of necessary protection, assume for example *that additional information* is stored for the purpose of re-identification of data subjects which happens only rarely and in exceptional cases. Consequently, the *additional information* is most of the time data at rest. At rest, protection can be as easy as encryption<sup>111</sup> or storing it on a USB memory stick and locking it away in the safe. To protect data in use at the same level

---

<sup>108</sup> See for example Art. 24(1) and 25(1) GDPR which both state that when implementing appropriate technical and organisational measures, the controller shall take into account "the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing".

<sup>109</sup> See for example Art. 32(1) GDPR that mandates the controller and processors to implement "appropriate technical and organisational measures to ensure a level of security appropriate to the risk".

<sup>110</sup> Data at rest can for example be protected through encryption.

<sup>111</sup> Encryption also adapts easily to varying requirements of the necessary protection. For example, assume that re-identification shall happen selectively, one data subject at the time. This can be supported by using a different encryption key per data subject. Similarly, assume that re-identification requires authorization by a dedicated (internal or external) authority. This can be done by giving this authority sole control over a master key that is necessary to yield the individual keys for data subjects.

of security is obviously much more onerous, requiring authentication, authorization of accesses, and confidentiality measures usually in a network environment.

#### 4.1.2 Risk reduction through pseudonymization

According to Recital 28, “[t]he application of pseudonymisation to personal data can reduce the risks to the data subjects concerned”. The following reasons how exactly pseudonymization reduces this risk.

For this purpose, it focuses on the essence of the separation. Unseparated (i.e., identified) data contain information about both, the “who” and the “what”. In essence, pseudonymization can be seen as a manner of processing that keeps the “who” and the “what” strictly separated. In particular, the *additional information* can be interpreted as information about “who”, while the *pseudonymized data* corresponds to the “what”.

Considering the fact that anonymous data are no longer subject to the GDPR<sup>112</sup>, it is evident that in absence of identification, the risk of processing for the rights and freedoms of natural persons is marginally low. While the *pseudonymized data* still permits identification, this is only possible with the use of *additional information*. In pseudonymization, access to *additional information* and thus identification of data subjects is proactively prevented. Therefore, if the implemented technical and organizational measures are effective, no identification of the *pseudonymous data* can take place. In other words, the “what” part of the data, in separation, represents only a marginal risk.

One may think that the risk inherent in the *additional information*, i.e., the “who” part, is equivalent to that of the identified data, i.e., the processing sans pseudonymization. But this may not be the case. Consider that the *additional information* predominantly informs about who is contained in the data set. It does not reveal any other information about those persons. From this point of view, the risk of the “who” part is usually significantly lower than that of the fully *identified data*.

Evidently, just knowing that a person is part of a data set may already represent sensitive information. For example, being contained in a data base of cancer treatments or in a registry of convicted felons obviously bears significant risks. On the other hand, the information that a person is a customer of a common online store or video streaming service may represent almost no risk at all.

Some kinds of identifying data elements may represent risks by in themselves. This is for example the case of credit card information or social security numbers. Apart from being identifying, it may also be possible to use it in ways that create harm or loss to a data subject. But even in these cases, the same risks are also present in the *identified data* and the latter usually bring additional risks through the identified inclusion of the “what” data.

This situation can be summarized as follows:

The risk represented by *identified data* is in general greater than the sum of the risks of *additional information* and *pseudonymized data* considered in separation.

$$\text{risk}(\textit{identified data}) \geq \text{risk}(\textit{additional information}) + \text{risk}(\textit{pseudonymized data})$$

---

<sup>112</sup> See Recital 26 GDPR.



Even if it is not possible to operate a processing activity completely as pseudonymization, partial use of pseudonymization can achieve significant risk reduction. This includes the following:

- If a processing can be broken up in multiple steps and some of those can be executed as a pseudonymization, it might be possible to significantly **reduce the recipients to whom identified data is disclosed**. Assume for example, that in an enterprise, the processing for quality assurance and for research do not require identified data. Then, pseudonymization can be used to avoid that personnel of the according departments get access to *identified data*.
- In some cases, even if re-identification is necessary after pseudonymization, the re-identification can be limited to special cases and thus affect **only a subset of data subjects**. For example, the processing of pseudonymized medical data may reveal cases where the data subject requires urgent medical attention. The re-identification, and thus use of identified data, can then be limited to the likely small number of affected data subjects.
- Even if a processing activity changes between pseudonymization and processing of identified data, the **temporal exposure** to a higher risk **may be significantly reduced**. Compared to exclusive processing of *identified data*, in the time spent on pseudonymization, the data bears a lower risk and is therefore less vulnerable.

#### 4.1.3 Importance of Pseudonymization in the GDPR

Pseudonymization is quite prominent in the GDPR. It was newly introduced in the sense that there was no mention of *pseudonymization* in the *European Data Protection Directive*<sup>113</sup> that preceded the GDPR.

*Pseudonymization* is defined in Art. 4(5) GDPR and some aspects are clarified by Recitals 26, 28 and 29. Recital 78 explicitly states that appropriate technical and organizational measures include “pseudonymising personal data as soon as possible”. *Pseudonymization* is probably the most prominent example for a technical and organizational measure mentioned in the GDPR. In addition to its definition, it is mentioned as an example of a measure (or safeguard) in five Articles:

- Art. 6(4)(e) in the context of compatible purposes,
- Art. 25(1) as an example for a measure in the context of data protection by design,
- Art. 32(1)(a) as an example of a security measure,
- Art. 40(2)(d) as an example of a possible concern of a code of conduct, and
- Art. 89(1) as an example of a safeguard in the context of further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.

In addition, Recitals 75 (on risks in general) and 85 (in regard of data breaches) explicitly mention the “unauthorised reversal of pseudonymisation” as a risk.

While the GDPR repeatedly also mentions also *encryption* as an example for a technical measure, compared to *pseudonymization*, this happens much less frequently (in three articles and one recital).

This prominence of *pseudonymization* in the GDPR underlines the importance that the legislator seems to have assigned to this concept.

---

<sup>113</sup> Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23.11.1995, p. 31–50, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31995L0046> (last visited 22/1/2021).

## 4.2 Pseudonymization in the GDPR

The definition of pseudonymization can be found in Art. 4(5) GDPR. The following discusses this definition and provides a technical interpretation.

The GDPR defined pseudonymization as follows:

Definition: **Pseudonymization** (according to Art. 4(5) GDPR)

“‘pseudonymisation’ means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;”

Some elements of this definition are the following:

- *Pseudonymization* is a **manner of processing**, i.e., not a transformation of a data set--which is a common technical interpretation of the concept. To capture this latter meaning, the term *data pseudonymization* will be introduced later.
- Evidently referring to the **pseudonymized data** that are being processed, the key requirement of pseudonymization is that “the personal data **can no longer be attributed to a specific data subject** without the use of additional information”.

This requirement is expressed in the context of processing. This processing comes with its own technical and organizational measures. The most important ones here are probably measures of **confidentiality** that limit disclosure of the data to the **intended recipients**. Confidentiality measures could be complemented by organizational measures that hold the recipients to a certain conduct (such as refraining from any attempt of identification).

Based on these considerations, the first half of the definition of pseudonymization in Art. 4(5) GDPR can be reworded as follows: **It shall not be possible for the intended recipients in the context of processing to identify data subjects in the pseudonymized data** except with the use of additional information.

- The second half of the definition of pseudonymization in Art. 4(5) GDPR is concerned with the additional information: “provided that such **additional information is kept separately** and is **subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person**”. The additional information thus needs to be kept separately and protected by technical and organizational measures. These **measures ensure**, i.e., **guarantee, that identification of data subjects cannot take place**.
- Considering that
  - (i) pseudonymized data do not permit the recipients to identify data subject except by using additional information and
  - (ii) that the measures protecting the additional information guarantee that identification cannot happen,

then, it follows that **identification cannot happen at all inside the realm of pseudonymization**.

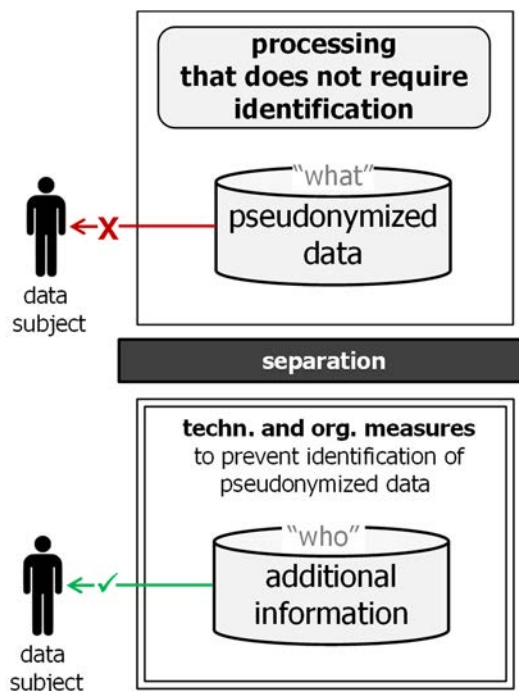
- Any re-identification must therefore take place outside of the realm of pseudonymization. This is for example the case for the processing of data subject requests that require data to be identified<sup>114</sup>. Re-identification and the processing of the request are then not part of pseudonymization. Much rather, they are part of a separate realm of processing of identified data.
- It is important for the understanding to explore what kinds of technical and organizational measures around the additional information can actually guarantee that identification does not take place. Considering that the coming together of pseudonymized data with the additional information permits re-identification of the data, it is clear, that these measures must prevent this. This can only be achieved by an impenetrable separation of pseudonymized data from the additional information. Any breach of this strict separation would contradict the (narrow) definition of *pseudonymization* given in Art. 4(5). This means for example that within the realm of pseudonymization, no party can get access to both, the pseudonymous data and the additional information.
- Note that the definition of pseudonymization clearly states that ***pseudonymized data is "personal data"***. This is further confirmed in Recital 26, 2<sup>nd</sup> sentence, that states: "Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person."

In addition to the definition of Art. 4(5) GDPR, Recital 29 provides additional help to interpret the concept of *pseudonymization*. It clarifies that pseudonymization (as a manner of processing) can be "possible within the same controller" as long as "that controller has taken technical and organisational measures necessary to ensure [...] that additional information for attributing the personal data to a specific data subject is kept separately". In other words, the additional information that permits identification of data subjects does not need to be kept by a (trusted) third party but can be managed by the controller itself. The controller is thus trusted to render undesired identification impossible. This requires the implementation of adequate measures.

---

<sup>114</sup> For example, the right of access (Art. 15 GDPR) necessarily bring together both, the (potentially pseudonymized) data and the full identity of the data subject.

The above interpretation of *pseudonymization* as given by the GDPR is visualized in Figure 12:



**Figure 12: Pseudonymization as a manner of processing according to Art. 4(5) GDPR.**

The upper part of the figure corresponds to the partial sentence “processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information” of Art. 4(5) GDPR.

The lower part corresponds to the partial sentence “provided that such additional information is kept separately”. It further visualized the partial sentence “[provided that such additional information] is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person”.

Note that in the lower part of the figure that is concerned with *additional information*, no processing is mentioned or implied in Art. 4(5). The additional information is solely kept for the possibility of exiting (through *re-identification*) or entering (through *data pseudonymization*) the realm of *pseudonymization*.

Between the upper and lower part of the figure respectively, a black bar represents the *separation*. Separation of the *pseudonymized data* from the *additional information* is a key concept of pseudonymization. It is this separation that guarantees “that the personal data can no longer be attributed to a specific data subject without the use of additional information”.

The figure also illustrates the technical and organizational measures to which the additional information is subject as a double box around the *additional information*. Since these measures “ensure that the personal data are not attributed to an identified or identifiable natural person”, they also enforce the mentioned separation. They are discussed in more detail in section 4.7 below.

### 4.3 The context of pseudonymization and access to additional information

Art. 4(5) chose to define *pseudonymization* in a very narrow manner. It is therefore useful to see it in its wider context. For this purpose, Figure 13 illustrates the situation.

In the middle of the Figure, the representation of *pseudonymization* from Figure 12 can be recognized. Its elements are grouped into a box that represents the realm of pseudonymization. There are two transformations that lead in and out of the realm of pseudonymization. Namely, these are *data pseudonymization* and *re-identification*. Both will be defined in more detail in section 4.5. These transformations bridge between the *realm of pseudonymization* and that of *processing of identified data*. Both of these transformations require access to both, the *pseudonymous data* and the *additional information*. In particular, *data pseudonymization* creates both by splitting *identified data* into a who and what part; and *re-identification* combines these two back into *identified data*.

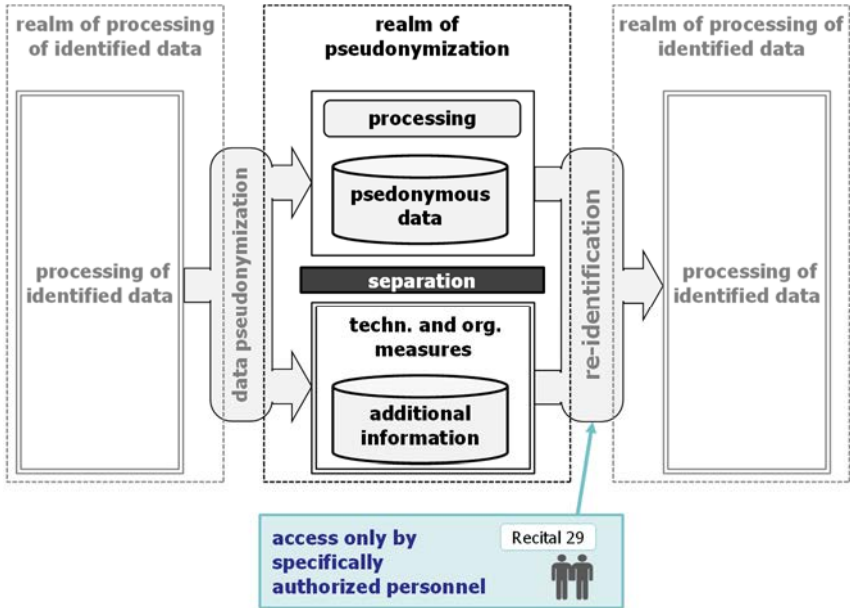


Figure 13: The context of pseudonymization.

In this scenario, the particularly protected access to *additional information* is the key to being able to transition between the realms of *pseudonymization* and *processing of identified data*, respectively. This key must be specifically guarded. For this reason, Recital 29 states in the context of additional information that “[t]he controller processing the personal data should indicate the **authorised persons** within the same controller.” It is particularly critical here to restrict who can exit the *realm of pseudonymization* and enter the *realm of identified data*. Therefore, persons who are granted access to both, the *pseudonymous data* and the *additional information* should be specifically authorized by the controller.

#### 4.4 Usage scenarios of pseudonymization

The previous section had shown how the narrow definition of *pseudonymization* provided by the GDPR is embedded in a wider context that includes the processing of identified data. The following gives some typical usage scenarios that illustrate this further.

Figure 14 shows a basic three-step scenario where the processing activity starts with the processing of *identified data*. Thereafter, the data is split into *pseudonymized data* and *additional information*. In that step, only the *pseudonymized data* is processed; the *additional information* is simply stored for later use. In a third step the *additional information* is used to re-identify the *pseudonymized data* for a final processing of *identified data*. The figure hints at the possibility that only a subset of the *pseudonymized data* may require re-identification. This scenario can obviously be generalized to more than three steps.

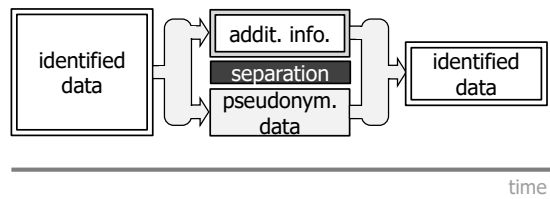


Figure 14: Pseudonymization with partial re-identification.

Figure 15 shows a simple scenario where after an initial step of processing *identified data*, only *pseudonymized data* are necessary and there is no need for re-identification. For this reason, the controller refrains from storing any *additional information*. This may for example apply to some cases of “further processing” for “compatible purposes” such as the “further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes” mentioned in Art 5(1)(b) GDPR. (See also Art. 6(4) and 89(1) GDPR).

In a variation of this scenario, the controller refrains from collecting any directly identifying data elements and the complete processing acts only on *pseudonymous data*.

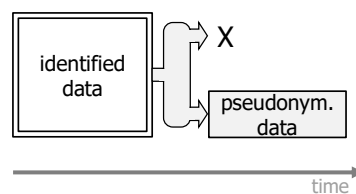


Figure 15: Pseudonymization without possibility of re-identification.

Figure 16 shows a scenario where the ongoing processing of *identified data* is accompanied by parallel pseudonymization. On reason to operate pseudonymization in this way is illustrated in the figure. In particular, it enables controllers to disclose only *pseudonymized data* for example to an external processor or internal department. The processor is then faced with a much lower risk. The controller keeps the *additional information* to feed the results of the processor’s work back into the processing of *identified data*. Evidently, this strategy avoids that fully identified data is disclosed to the processor and its employees. This reduces the overall risk of the processing activity.

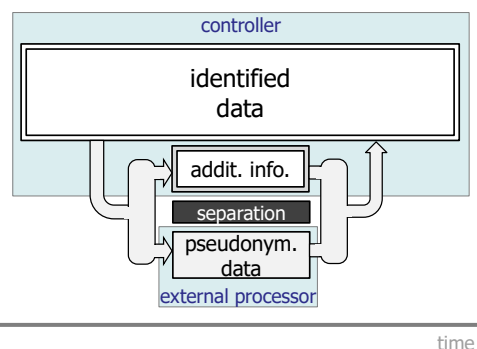


Figure 16: Pseudonymization in support of (for example) simplified outsourcing.

Figure 17 shows a scenario of continuous data acquisition. In particular, incrementally, at a later point of time, i.e., after the initial data acquisition, additional data for existing data subjects arise. For example, this may be the case for the research on long-lasting medical treatments.

When additional *identified data* arises for a data subject that is already known, the already stored *additional information* is necessary to use the same *pseudonym*<sup>115</sup> that was used earlier for this data subject. In this manner, the pseudonymous data that refers to the same data subject can be linked.

Note that in this scenario, the *additional information* is needed even if re-identification of the pseudonymized data is unnecessary. (The *additional information* can also be *one-directional*; see definition below).

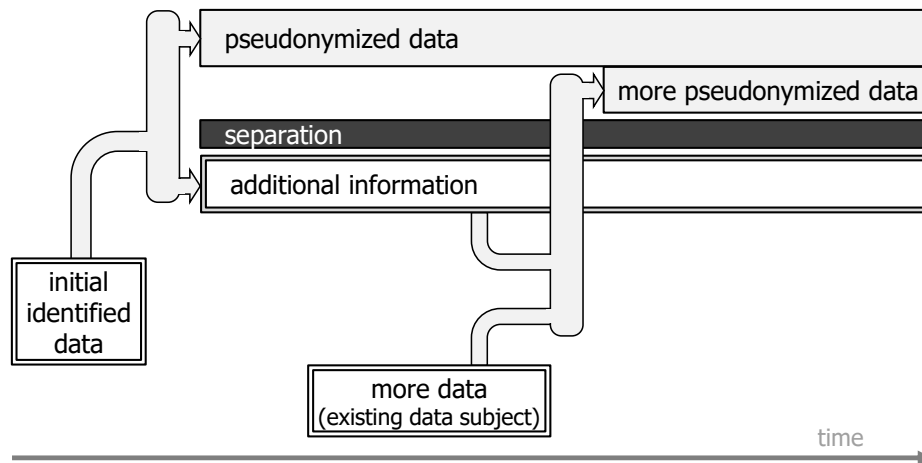


Figure 17: Pseudonymization of data set that grows over time.

## 4.5 Definition of concepts relevant to pseudonymization

*Pseudonymization* is defined in Art. 4(5) GDPR as a manner of processing. The GDPR lacks definitions for closely related concepts, however, including the following:

- Data pseudonymization,
- re-identification,
- identified data,
- pseudonymous data,
- additional information, and
- pseudonym.

The following attempts to provide precise, mutually consistent definitions of these concepts. In addition, it distinguishes different kinds of additional information.

### 4.5.1 Data pseudonymization

The GDPR implies that *identified data* can be separated into *pseudonymized data* and *additional information*. In the technical literature, this transformation is often referred to simply as *pseudonymization*. This term is already used in the GDPR with the semantics of “a manner of processing”, however. Therefore, the term *pseudonymization* is used in its meaning provided by the GDPR. To distinguish the different concepts, the transformation is then called *data pseudonymization*. In the case where data collection was not already limited to pseudonymous data, *data pseudonymization* is a pre-requisite for *pseudonymization*.

<sup>115</sup> Here, the generally known term of *pseudonym* is used. It would be more precise to use the term *pseudonymous handle* that is a special kind of pseudonym that will be defined later.

Definition: **Data pseudonymization**

*Data pseudonymization* is a transformation that takes *identified data* as input and creates two output data sets, namely *pseudonymous data* and *additional information*, respectively.

Data pseudonymization is illustrated in Figure 18.

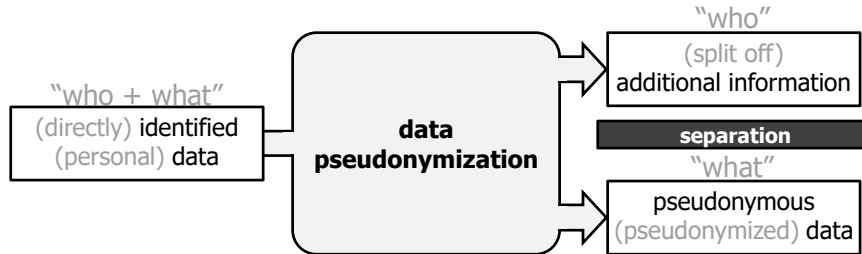


Figure 18: Data pseudonymization.

## 4.5.2 Re-identification

The following defines two concepts of re-identification. Namely it first defines the term in a general context and then in the specific context of (data) pseudonymization.

### 4.5.2.1 (General) re-identification

The GDPR also states that *pseudonymous data* can be attributed to specific data subjects using *additional information*. This is the inverse transformation of *data pseudonymization*. It is called *re-identification* and illustrated in Figure 19. Note that the notion of *additional information* in re-identification is *general*. It is not limited to that *additional information* that the controller keeps separately from the *pseudonymized data*. It could be any *additional information* existing anywhere (also outside of the controller) as long as it is suitable to identify data subjects in the *pseudonymized data*.

Definition: **(General) re-identification**

*Re-identification* in the general sense is a transformation that takes *pseudonymous data* and *additional information* as input and creates *identified data* as output.

The concept is general and de-coupled from *pseudonymization* in the sense that the *additional information* is not limited to that resulting from *data pseudonymization* and stored by the controller. Much rather, **any** *additional information* can be used, including and most commonly that existing outside of the controller.

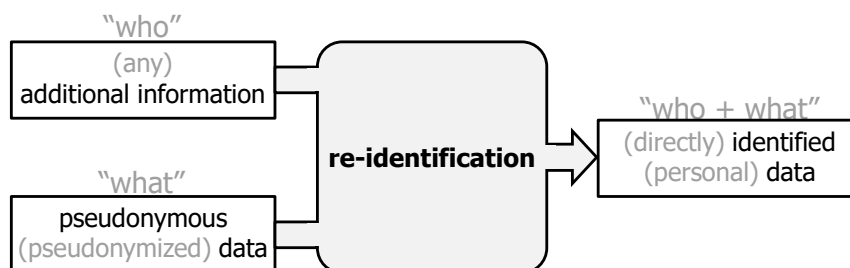


Figure 19: General re-identification.



#### 4.5.2.2 Planned re-identification

It is often practical to distinguish *re-identification* in general from that executed by the controller in the context of pseudonymization. For this purpose, a distinct special case of the concept is introduced and called *planned re-identification*. In general, the controller stores the (*split-off*)<sup>116</sup> *additional information* for the purpose of being able to re-identify the *pseudonymized data* at a later point in time<sup>117</sup>. Since this kind of *re-identification* is thus foreseen and planned, this specific kind of *re-identification* is therefore called *planned re-identification*. It is defined here and illustrated in Figure 20.

**Definition: *Planned re-identification***

*Planned re-identification* is the special case of *re-identification* where the *additional information* is that resulting from *data pseudonymization* and stored by the controller.

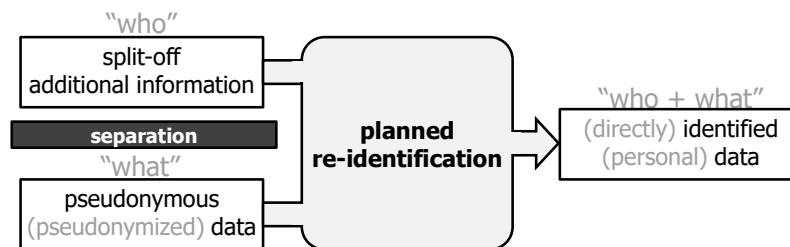


Figure 20: Planned re-identification.

#### 4.5.3 (Directly) identified (personal) data

Data processing sans pseudonymization usually operates on *identified data*. They are defined as follows:

**Definition: (*Directly*) identified<sup>118</sup> (personal) data**

*Directly identified personal data*, or more shortly *identified data*, is personal data that allows direct identification of data subjects.

This is for example the case when the data includes names or commonly used unique handles. The term is synonym to the expression “personal data relating to an *identified* data subject”. It implies that the data can be *directly linked* to information assets in possession of the actor who identifies (see Figure 5: Identification of a data subject.” above).

<sup>116</sup> The term *split-off additional information* will be defined below to denote the *additional information* stored by the controller.

<sup>117</sup> Note that Art. 11 GDPR states that additional information shall not be stored for the sole purpose of complying with the requirements of the GDPR, such as the implementation of data subject rights. Data minimization (Art. 5(1)(c) GDPR prohibits to store the additional information in the case that it is not required for the purposes of processing. The need for additional information most commonly arises for re-identification. The one exception is to incrementally add to a set of pseudonymized data as was illustrated in the usage scenario around Figure 17 above.

<sup>118</sup> The term “identified” seems a good description of the essence since the data contains both, the “who” and the “what”; if it contained only the “who”, “identifying” would likely be a better choice for the concept.

Using the definition of *identification* and its visualization in Figure 5 from above, identified personal data is illustrated in Figure 21.

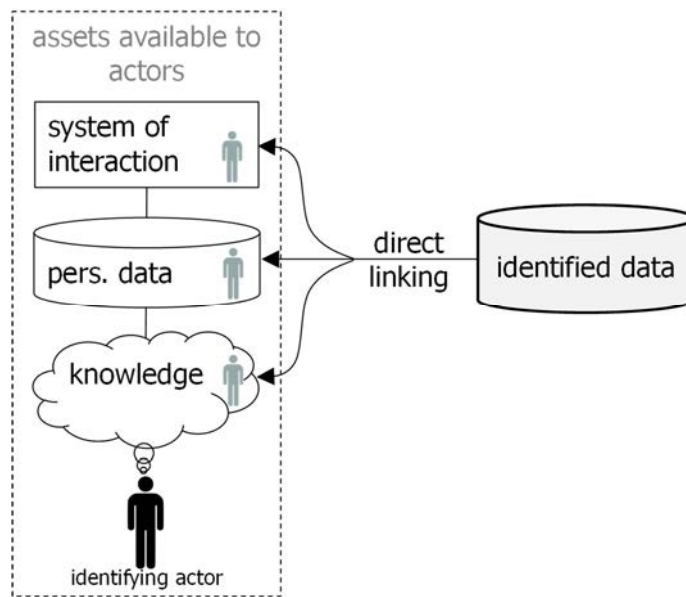


Figure 21: Identified data.

#### 4.5.4 Pseudonymous data

The following provides two definitions for *pseudonymous data*. The first attempts to capture the general use of the term in a wider sense; the second uses a narrower definition that imposes stricter requirements that stem from the definition of *pseudonymization* in the GDPR.

##### 4.5.4.1 (General) pseudonymous data

In a wider, more general sense, data is often called *pseudonymous* if it refrains from including directly identifying data elements (such as names or unique handles) to refer to data subjects. This common use of the term is captured in the following definition.

**Definition: (General) pseudonymous data**

*General pseudonymous data*, or simply *pseudonymous data*, is data that refrains from containing any directly identifying data elements (“identifiers”) such as names, commonly used unique handles, or common quasi-identifiers.

##### 4.5.4.2 Strictly pseudonymous data

*General pseudonymous data* fails to satisfy the requirements implied in Art. 4(5) GDPR. In particular, Art. 4(5) states that “the personal data can no longer be attributed to a specific data subject without the use of additional information”. Simply leaving away knowingly identifying data elements does not guarantee this. In particular, to capture the essence of Art. 4(5), the general definition above, has at least two shortcomings:

- Even if no knowingly identifying data elements are present, persons can still be identified (“recognized) by unique attribute values and combinations of attribute values (e.g., “the old, red haired guy in the yoga class”).

- General pseudonymity fails to define a precise context in which the inability to identify can be evaluated. In particular, the context should define the actors who can potentially identify data subjects (and thus the available assets) and the technical and organizational measures that are in place to prevent identification.

In absence of a context and thus technical and organizational measures (incl. access control), pseudonymous data are accessible by anyone. It is probably impossible to know what assets and additional information about data subjects is available to different actors out there. It is therefore also probably impossible to state that none of these can identify data subjects in the pseudonymous data (as is required by Art. 4(5) GDPR).

To mend these shortcomings, a *stricter* definition of *pseudonymous data* is given. As a preparation of this definition, Art. 4(5) must first be analyzed in more detail.

The personal data that is processed during *pseudonymization* will be called *strictly pseudonymous data*. This refers to Art. 4(5) GDPR that uses the wording “processing of **personal data**<sup>119</sup> in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information”.

The cited wording of Art. 4(5) GDPR contains two aspects:

- (i) The *pseudonymous data* can no longer be attributed to a specific data subject without the use of additional information, and
- (ii) This requirement is formulated in the context of the processing that makes up *pseudonymization*.

To better understand this, the model of *identification* represented in Figure 5 above is adapted here. The application of this model to *pseudonymization* is illustrated in Figure 22.

The model translates “attribution to a specific data subject” of (i) into “direct linking to information assets available to the identifying actors”.

Art. 4(5) expresses the requirement (i) in the well-defined context of the processing of *strictly pseudonymous data* (ii). This context defines the protective technical and organizational measures that are implemented by the controller for this processing. They cover at least the following aspects:

- Restriction of the access<sup>120</sup> to the pseudonymous data to **intended recipients**<sup>121</sup>,
- prevention of these recipients to access the **internal additional information**,
- prevention of these recipients to combine potential **external additional information** with the pseudonymous data.

In presence of these protective measures, direct linking of the pseudonymous data to the assets available to the intended recipients shall be impossible.

This means that the concept of *strictly pseudonymous data* can only be used in a specific context of processing. *Strictly pseudonymous data* taken outside of this context can well be directly linkable to available (information) assets, for example because:

- other actors than the intended recipients may have access to assets that permit direct linking, or

---

<sup>119</sup> Highlighting added by the author.

<sup>120</sup> Note that since pseudonymous data are still personal data, this is mandated by Art. 5(1)(f) “integrity and confidentiality” GDPR.

<sup>121</sup> According to Art. 4(9) GDPR, “recipient’ means a natural or legal person, public authority, agency or another body, to which the personal data are disclosed, whether a third party or not”.

- indirect linking is still possible in absence of adequate protective technical and organizational measures that prevent access to (internal or external) *additional information*.

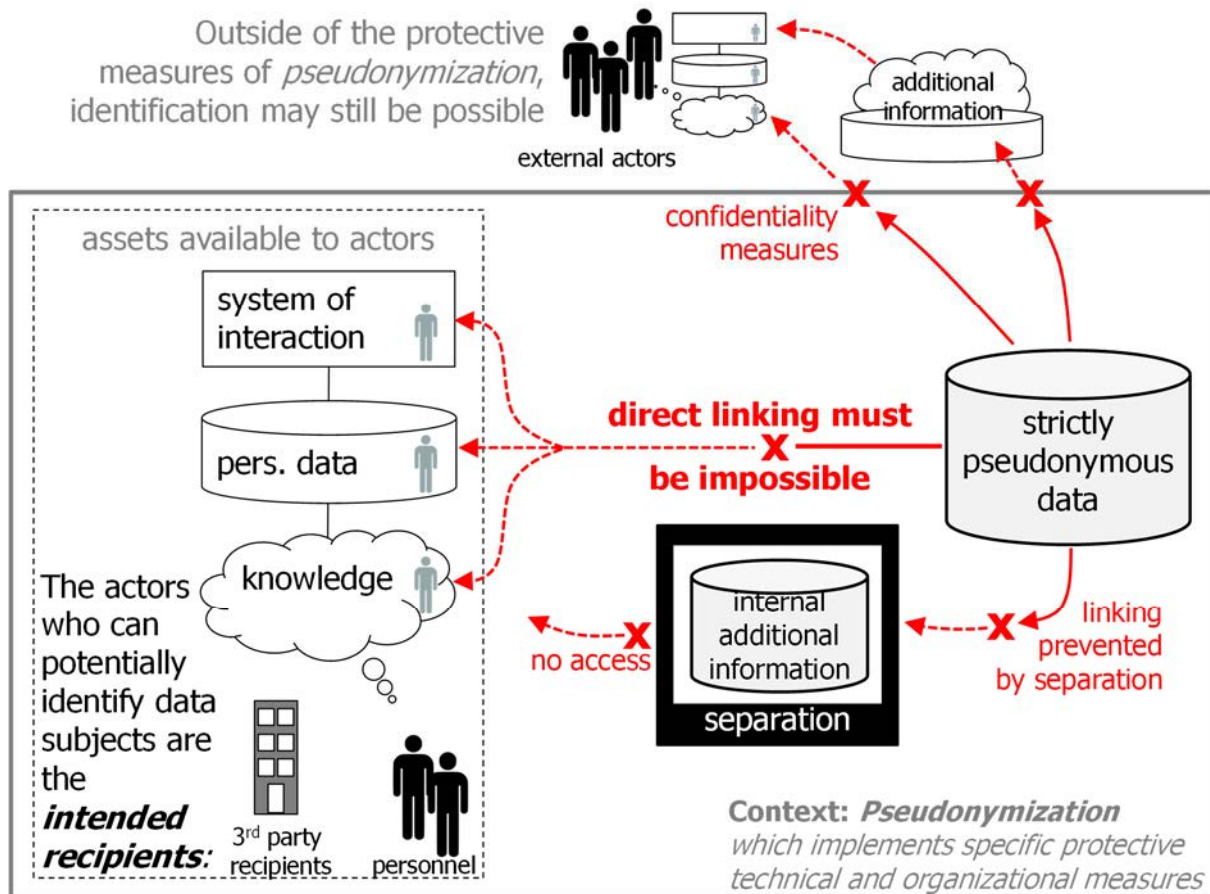


Figure 22: In the context of pseudonymization, intended recipients are unable to identify data subjects in strictly pseudonymous data.

This analysis is summarized in the following definition:

**Definition: Strictly pseudonymous data**

Data is *strictly pseudonymous* in the context of *pseudonymization*, if, in presence of the technical and organizational measures of the *pseudonymization*, the *intended recipients* are unable to *directly identify* data subjects. In absence of these measures, indirect identification using *additional information* is still possible. *Strictly pseudonymous data* is a special case of (*general*) *pseudonymous data* that satisfies the stricter requirements implied by Art. 4(5) GDPR.

Note that this text predominantly discusses *strictly pseudonymous data*. Being a special case of (*general*) *pseudonymous data*, it is still correct to call them simply *pseudonymous data*. This has been done excessively in this text. It also applies to the labels of *pseudonymous data* in many figures above. When the simplified version of the concept is used, it should be clear from the context provided by the text, that the discussion is concerned with *strictly pseudonymous data*. This is basically always the case in this text, unless where it is explicitly stated that it deals with *general pseudonymous data*.

#### 4.5.4.3 Pseudonymized data

*Strictly pseudonymous data* is the data processed during *pseudonymization*. It is often created as an output of *data pseudonymization*. To specifically refer to *strictly pseudonymous data* that is created in this manner, the term *pseudonymized data* can be used.

**Definition: *Pseudonymized data***

*Pseudonymized data* is strictly *pseudonymous data* that is created as an output of *data pseudonymization*.

Note that *strictly pseudonymous data* is not always the result of *data pseudonymization*. For example, data can be collected in a manner such that it is already *strictly pseudonymous*. This includes for example to refrain from collecting directly identifying data elements and manage potentially unique attribute values. For this reason, *pseudonymized data* is not used exclusively, but the more general concept of *strictly pseudonymous data* is still necessary.

#### 4.5.5 Additional information

This section defines a more general, wider and a more specific, narrower concept of *additional information*.

The general definition describes any information anywhere that is suitable to be used to identify data subjects in pseudonymous data; the more specialized definition, called *split-off additional information*, refers to that *additional information* which is kept separately by the controller. The latter is typically the results of *data pseudonymization* and is used for *planned re-identification*.

In addition, this section defines different possible technical formats (called types) used to represent *additional information*. This distinction is later on used to capture different levels at which controllers are able to re-identify *pseudonymous data*.

##### 4.5.5.1 (General) additional information

*Additional information* is a concept that is central to pseudonymization. The following provides first a general, wider definition of *additional information*.

**Definition: (General) additional information**

*Additional information* is knowledge or data that can be used for *indirect identification* of at least one data subject in *pseudonymous data*. For that purpose, the additional information must establish a relation between

- (i) directly identifying data elements that relate to *identified* data subjects and
- (ii) information elements that permit direct linking to the *pseudonymous data*.

The latter linking can be based on

- *unique handles* (including *pseudonyms*) as well as
- (single or combinations of) unique values, quasi-identifiers, or identity-relevant properties.

The general concept of *additional information* is independent of *data pseudonymization*. While one of the outputs of *data pseudonymization* is indeed (*split-off*) *additional information*, *additional information* can also exist independently and be held by other parties than the controller. Any data anywhere that permits (at least partial) identification of the *pseudonymous data* at hand is therefore considered to be *additional information*.

Figure 23 illustrates how *(general) additional information* establishes a relation between data elements that uniquely match to the *pseudonymous data* on one end, and data elements that uniquely identify data subjects on the other. The figure also provided examples for such data elements.

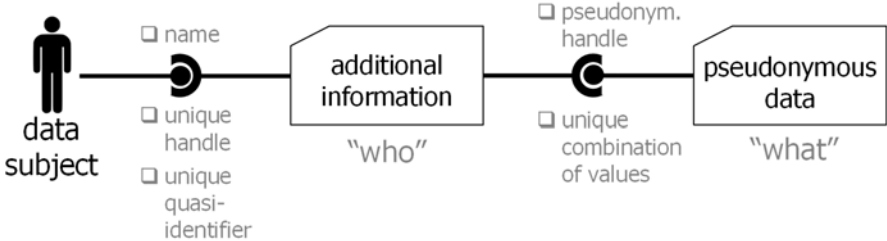


Figure 23: Additional information links pseudonymous data to a data subject.

#### 4.5.5.2 Split-off additional information

While the above definition of *additional information* is general, *split-off additional information* is a specialized form that addresses the *additional information* typically used in *pseudonymization*.

Definition: **Split-off additional information**

*Split-off additional information* is the *additional information* that results from *data pseudonymization*.

Since it is designed to re-identify the pseudonymized data, on one side of the relation, it typically uses the *pseudonym* (or more precisely, the *pseudonymous handle*) to link to the (*strictly*) *pseudonymous data*. (This contrasts the general concept of *additional information* where such linking can also be based on unique combinations of values). On the other side of the relation, it typically uses a *unique handle* that is in use by the controller (such as a customer ID) to identify data subjects.

While *additional information* in general identifies at least one data subject in the set of *pseudonymous data*, *split-off additional information* usually identifies all data subjects in the set of *strictly pseudonymous data*.

#### 4.5.5.3 Different types of additional information

While the concept of *additional information* was defined above, the following distinguishes different types of technical representations of additional information. In particular, it distinguishes

- lookup-based and
- formula-based

*additional information*, as well as

- one-directional and
- bi-directional

*additional information*. These two distinctions are independent and can be combined.

The distinctions are particularly useful for *split-off additional information*. They are helpful to reason about data minimization and about the possibility a controller has to re-identify data.

The different types are as follows:

Definition: **Lookup-based additional information**

*Lookup-based additional information* takes the form of a lookup table where every row, pertaining to a single data subject, contains both, (one or several) directly identifying data elements and (one or several) data elements that permit linking to the pseudonymous data. The simplest form of *lookup-based additional information* consists of one column with a *unique handle* for data subjects and one with a *pseudonym* (i.e., *pseudonymous handle*, see below). *Lookup-based additional information* is always *bi-directional* (see definition below).

Figure 24 gives an example for *lookup-based additional information*.

<b>unique handle 1</b>	<b>pseudon. handle 1</b>
<b>unique handle 2</b>	<b>pseudon. handle 2</b>
<b>unique handle 3</b>	<b>pseudon. handle 3</b>
...	...
<b>unique handle n</b>	<b>pseudon. handle n</b>

lookup table

**Figure 24: Lookup-based additional information.**

Definition: **Formula-based additional information**

*Formula-based additional information* takes the form of a function expressed by a formula whose input consists of (one or several) directly identifying data elements and whose output are (one or several) data elements that permit linking to the pseudonymous data. The simplest form of *formula-based additional information* takes a *unique handle* of data subjects as input and yields a *pseudonym* (i.e., *pseudonymous handle*, see below) as output.

Note that an inverse function may or may not exist. In the example where the function is an encryption, the inverse function exists in the form of decryption. In the example where the function is a cryptographic one-way function (such as an HMAC), the inverse function does not exist.

Note that in order to prevent linking between *identified data* and *pseudonymous data*, the function used in formula-based additional information should include a **secret**. In particular, identification was defined above as the possibility to link *pseudonymous data* with a data subject, independently of the direction of such linking. So *identification* includes both, locating the person that belongs to the data, and vice versa. Functions without a secret prevent locating the person who belongs to the data; but they allow to compute the *pseudonym* from a known person handle. This means that it is still possible to link between *identified data* and *pseudonymous data*. For this reason, known or easy to guess cryptographic one-way functions (such as cryptographic hashes or digests<sup>122</sup>) are in general unsuited for pseudonym creation. Much rather, the function should contain a secret<sup>123</sup> as is the case in the example of a keyed message authentication code<sup>124</sup> (HMAC). This is also in line with Art. 4(5) GDPR which implies that *additional information* needs to be protected with adequate technical and

---

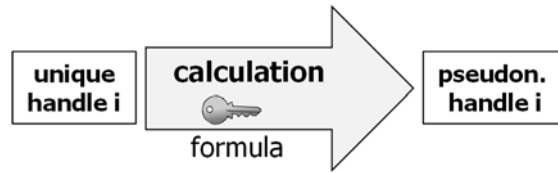
<sup>122</sup> See [https://en.wikipedia.org/wiki/Cryptographic\\_hash\\_function](https://en.wikipedia.org/wiki/Cryptographic_hash_function) (last visited 18/2/2021).

<sup>123</sup> Note that a “salt” is not always considered to be a secret (see for example <https://crackstation.net/hashing-security.htm> and <https://security.stackexchange.com/questions/131731/why-can-salts-be-public>) but a measure to render attacks with rainbow tables computationally more expensive. This stands in contrast to “pepper” that is sometimes also called “secret salt” (see [https://en.wikipedia.org/wiki/Pepper\\_\(cryptography\)](https://en.wikipedia.org/wiki/Pepper_(cryptography))).

<sup>124</sup> See <https://en.wikipedia.org/wiki/HMAC> (last visited 18/2/2021).

organizational measures. It would be futile to protect anything that is publicly known or can be easily guessed.

Figure 25 illustrates an example of *formula-based additional information*.



**Figure 25: Formula-based additional information.**

Additional information belongs to one of the two above types. Independently of this distinction, another independent distinction can be made:

**Definition: *Bi-directional additional information***

*Bi-directional additional information* permits to use the *additional information* to link in both directions:

- From a given record in the *pseudonymous data* to the data subject, and
- from a known data subject to the corresponding record in the *pseudonymous data*.

*Lookup-based additional information* and encryption (i.e., an example of *formula-based additional information*) are examples for *bi-directional additional information*.

**Definition: *One-directional additional information***

*One-directional additional information* permits to use the *additional information* only in one direction:

- From a known data subject to the corresponding record in the *pseudonymous data*.

A typical example of *one-directional additional information* is a one-way function (such as a keyed HMAC). It usually maps a directly identifying *unique handle* of the data subject into a *pseudonym* (i.e., *pseudonymous handle*, see below) that can be linked to the *pseudonymous data*. Since a one-way function fails to have an inverse, it is not possible to inversely compute the *unique handle* of the data subject from the *pseudonym*.



Figure 26 further illustrates the different types of *additional information* by providing common examples.




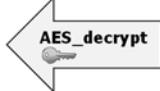

type of additional information	forward function	inverse function
	identified $\Rightarrow$ pseudonymous	pseudonymous $\Rightarrow$ identified
lookup-based bi-directional	 lookup table	 lookup table
formula-based bi-directional		
formula-based one-directional		X

Figure 26: Examples of different types of additional information.

#### 4.5.6 Pseudonyms

In the literature, the term of *pseudonym* is used with a variety of semantics for data elements that refer to a person without directly revealing the person’s identity. *Pseudonym* has thus used to refer to concepts that include self-chosen nicknames of persons (“superman99”), IP addresses, *pseudonymous handles* (see definition below), as well as group-, role-, and transaction-pseudonyms.

To embrace this general use of the term *pseudonym*, the following first provides a general, wide definition of the concept. Then, to cater to the specific context of *pseudonymization*, a special case of a *general pseudonym*, called *pseudonymous handle*, is defined in a more specific, narrower sense.

##### 4.5.6.1 (General) pseudonyms

The following provides a wide definition of the term:

Definition: **(General) pseudonym**

A *general pseudonym* or simply *pseudonym* is a data element that refer to a person without directly revealing the person’s identity.

The following defines a specialized kind of *general pseudonym* suited for pseudonymization:

##### 4.5.6.2 Pseudonymous handles

The more specialized, narrower definition of *pseudonymous handle* refers to those *pseudonyms* that are typically contained in *split-off additional information* and *strictly pseudonymous data*. *Pseudonymous handles* are typically created as part of *data pseudonymization*.

Definition: **Pseudonymous handle**

A *pseudonymous handle* is a *unique handle* created in a separate *identity domain* with the sole purpose of creating a relation between *split-off additional information* and *strictly pseudonymous data*. This relation is established by inserting the *pseudonymous handle* in both, the *split-off additional information* and the *strictly pseudonymous data*. This enables easy *deterministic linking* based on equality matching.

Since the *pseudonymous handle’s* identity domain is separate, it is impossible to link the *pseudonymous handle* to any other data sets but the *strictly pseudonymous data* and the *split-off additional information*.

Note that technically, a *pseudonymous handle* is also a (*general*) *pseudonym*. Therefore, where it is clear from the context that the text is concerned with a *pseudonymous handle*, it can be simply referred to as *pseudonym*. With the exception of the above definition of *general pseudonym*, the present text is exclusively concerned with *pseudonymous handles*.

## 4.6 Data pseudonymization in detail

The following describes a typical procedure of how to perform *data pseudonymization*. In other words, it describes the steps of how to construct a tuple of *strictly pseudonymous data* and *split-off additional information* starting from *identified data*. It depicts the common case where the *identified data* was previously used for other purposes. Pseudonymization could then constitute “further processing” (see Art. 5(1)(b) and 89(1) GDPR) that pursues its own purposes.

The overall procedure of *data pseudonymization* is illustrated in Figure 27 and discussed in the following.

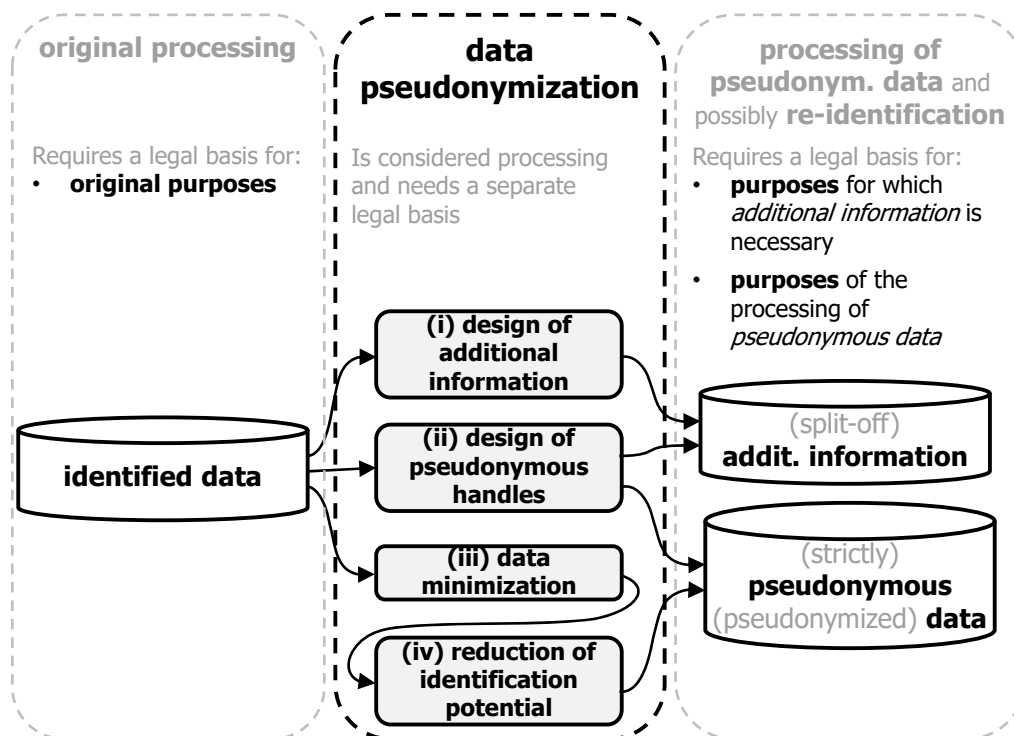


Figure 27: Functional details of data pseudonymization.

**Preparatory step:** In preparatory step, controllers need to specify the **purposes** pursued by the *pseudonymization*, i.e., the processing after *data pseudonymization*. This includes both,

- the purposes for keeping *(split-off) additional information* and
- the purposes pursued by the processing of *(strictly) pseudonymous data*.

Clarity about these purposes is important to guide several processing steps of *data pseudonymization*.

This is most evident in the *data minimization*<sup>125</sup> step (iii), since it filters out all data and detail that is unnecessary to fulfill the stated new purposes.

<sup>125</sup> Note that data minimization is one of the principles of data protection (see Art. 5(1)(c) GDPR).

It is similarly crucial to the step of *design of additional information* (i). More precisely, also this step can be seen as a variation of data minimization: Identifying data elements within the *additional information* can be kept only if they are necessary for legitimate purposes. A precise specification of these purposes is therefore an important input into the data pseudonymization procedure. This will be explained further below.

In the sequel, the actual processing steps that constitute *data pseudonymization* are described:

**(i) Design of additional information:** Controllers have to make certain design-decisions about the additional information. This task is guided by the purposes for which *additional information* is necessary in the first place.

(a) A first decision that controllers need to make is whether their **purposes require the storage of additional information at all**. This is most often equivalent with the question of whether *re-identification* of data subjects is necessary. Another reason for which *additional information* is necessary is to handle incremental growth of personal data that affects already existing data subjects (see usage scenario of Figure 17 above).

If *additional information* is unnecessary for the purposes, *data minimization* and *storage limitation* (see Art. 5(1)(c) and (e) GDPR) mandate that no *additional information* be stored (see usage scenario of Figure 15 above). Note that Art. 11 GDPR states that it is not necessary to store additional information for the sole purpose of complying with requirements of the GDPR, such as the implementation of data subject rights.

(b) Once established that additional information is indeed necessary, controllers need to decide whether it has to be **one- or bi-directional**. When re-identification is necessary, the *additional information* must always be *bi-directional*. When an incremental growth of personal data has to be handled, it is sufficient that the *additional information* is *one-directional*. *Data minimization* and *storage limitation* (see Art. 5(1)(c) and (e) GDPR) mandate that *one-directional* much rather than *bi-directional additional information* shall be used if it is sufficient for the purposes.

(c) One further decision is which **directly identifying data elements** shall be used for the *additional information*.

Assume for example, that the *additional information* shall be used in rare cases to re-identify data subjects in order to contact them. This may for example be the case when processing pseudonymous health data that may reveal that a specific data subject suffers from certain medical conditions that require rapid medical attention or intervention. The controller then needs the *additional information* in support of the purpose of contacting the affected data subjects. Consequently, the identifying data elements should be those suited to establish such contact (such as a telephone number or e-mail address). This example corresponds to the scenario expressed in Figure 14 above.

In another example, assume that an external processor received pseudonymous data for analysis and that the result of the analysis has then be re-identified by the controller for further processing. This example corresponds to the scenario expressed in Figure 16 above. In this case, the identifying data element should be the unique handle that is used in the processing of the *identified data*.

**(ii) Design of pseudonymous handles:** This step affects both, the *split-off additional information* and the *strictly pseudonymous data* since *pseudonymous handles* are part of both. The decision to make here is how to actually create the pseudonymous handles. A definition of the concept of *pseudonymous handles* was given in the previous section; different methods for creating pseudonyms were discussed in section 3.7.5.1 above. In summary, pseudonyms can be created independently (e.g., as random numbers) or derived from certain identifying data elements (e.g., by using a cryptographic one-way function or encryption). The present step of *data pseudonymization* decides which is the most suitable method to use.

**(iii) Data minimization:** The *identified data* were designed to support a set of original purposes. The processing step of *data minimization* eliminates all data that are no longer necessary for the new purposes pursued by the processing of the *strictly pseudonymous data*. This could entail both, the elimination of complete data elements, or the reduction of detail through generalization. An example for the latter is to generalize a precise locations (represented by latitude and longitude coordinates) to larger areas (such as a ZIP area or a county).

Note that while functionally, *data minimization* may be indistinguishable from the *reduction of identification potential* (i.e., step (iv), see below), they are conceptionally distinct: the former reduces information content since it is no longer necessary to fulfill the new purposes; the latter may use the same transformations in order to prevent direct identification of individuals through the linking of data. *Data minimization* is listed here explicitly since certain data elements may be free of any risk of linking, but anyhow have to be removed during *data pseudonymization*.

**(iv) Reduction of identification potential:** The *strictly pseudonymous data* are constructed by reducing the identification potential of the *identified data*. This is achieved by applying appropriate transformations to reduce the identification potential of the *identified data* set until the resulting data cease to permit the direct identification by the intended recipients (see definition of *strictly pseudonymous data* above).

Section 3.7.5 above has provided an overview of transformations that reduce the identification potential. In summary, the most important are possibly deletion, generalization, slicing to reduce the dimensionality, and noise injection. These belong to both, the category of

- truthful transformations which reduce the level of detail in the data and
- transformations that introduce deviations from the truth (i.e., errors).

Some typical examples of transformations used during data pseudonymization shall illustrate the concept:

- Typically, all **unique handles** must be deleted<sup>126</sup>.
- **Quasi-identifiers** that permit direct recognition of persons must be either generalized or deleted.
- **Unique values** and **unique combinations of identity-relevant properties** have to be transformed with methods such as generalization, error injection, top-coding, or deletion.

As was illustrated above in section 3.7.5.4, these transformations gradually reduce the identification potential of the data. In particular, they gradually delete more data elements, reduce the level of detail contained in the data, or add noise (i.e., error) to impede linking. So the key question is how much identification potential needs to be reduced until direct identification is no longer possible.

As follows from the definition of *strictly pseudonymous data*, this question can be answered in the well-defined context of the *pseudonymization* at hand, including its technical and organizational measures and its intended (internal or external) recipients of the *strictly pseudonymous data*.

---

<sup>126</sup> Note that the *pseudonymous handle* is not present in the *identified data* but only created during *data pseudonymization*.

Once the recipients are identified, controllers need to assess what information assets are reasonably likely<sup>127</sup> available to them. These information assets can include the following:

- **Other data** kept by the controller for other processing activities that is also accessible<sup>128</sup> to the personnel with access to the *strictly pseudonymous data* at hand,
- possible **knowledge** about data subjects **in the head of personnel** (for example when they process data pertaining to close acquaintances), and
- **external data that is readily available**<sup>129</sup> to the personnel (as for example data that can easily be looked up on the Internet from the work place).

The question of whether the identification potential is reduced sufficiently to reach strict pseudonymity now boils down to whether the available *identified* information assets can be linked to the *strictly pseudonymous data*. Having identified these information assets and knowing the content of the *strictly pseudonymous data*, this becomes a well-defined task<sup>130</sup>. Since only the linking methodology that is reasonably likely used<sup>131</sup> by the known actors (i.e., intended recipients) has to be considered, complex linking methods can often be excluded. Organizational measures that prohibit<sup>132</sup> personnel to attempt any linking may further exclude possibilities of identification.

## 4.7 Technical and organizational measures for pseudonymization

The following provides more detail on technical and organizational measures that a controller can consider to implement in the context of *pseudonymization*. It focuses on both, (i) measures to which the *split-off additional information* is subjected and that enforce the required separation and (ii) measures to prevent direct-identification of the *strictly pseudonymous data*.

(i) **Measures to protect the additional information:**

The following lists measures that implement the separation of *split-off additional information* from the processing of the *strictly pseudonymous data*. The *additional information* is necessary to re-identify the *pseudonymous data* and thus to exit the realm of *pseudonymization*. The following measures prevent or control such an exit.

- Technical measures such as **encryption** of *additional information*, when it is data at rest, or **access control**, when it is data in use, are obviously necessary measures. Access control includes authentication, authorization and logging of access (creating an audit trail).

---

<sup>127</sup> The term “reasonably likely” is used on Recital 26, sentence 3, GDPR in a comparable context. The assessment of available assets must take the implemented technical and organizational measures into account.

<sup>128</sup> In case such other data exists but is not accessible to the personnel working with the pseudonymous data, the controller must obviously be appropriate technical and organizational measures to deny such access.

<sup>129</sup> While this data is certainly physically external and could therefore be considered to be “additional information”, it seem reasonable to include this data. After all, its access may be possible from the work place and may be seamless and indistinguishable from the access of local data.

<sup>130</sup> In particular, the task of determining whether data is indeed *strictly pseudonymous* is easy in comparison of determining whether data is *anonymous* (see below). This is due to the fact that the former determination is made in a very well-defined context, while the latter must consider any (realistically) possible context and thus introduces significant uncertainty.

<sup>131</sup> See Recital 26, sentence 3, GDPR.

<sup>132</sup> This can for example be achieved through a contractual agreement and reinforced through training.

- As recommended in Recital 29 GDPR, the controller should **explicitly authorize the personnel** who have access to the *split-off additional information* and can thus exit the realm of *pseudonymization*. It is good practice to **document** such authorizations and to **keep them up to date** following fluctuations in personnel.
- The **conditions** under which access to the *split-off additional information* (and thus re-identification) is authorized by the controller shall be **explicitly specified and documented**.
- The **procedures** to be followed when accessing *split-off additional information* for re-identification could be **authorized and documented** by the controller. Such a procedure can for example ascertain that all the access conditions have been verified and that access is properly authorized.

Since the access to *split-off additional information* is typically the key to re-identification, a more comprehensive procedure that captures the complete re-identification could be defined. In addition to accessing *split-off additional information*, in such a procedure also *strictly pseudonymous data* has to be accessed. The procedure could then, for example, minimize the re-identified data by restricting the used *additional information* to that of a single data subject and limiting the associated *pseudonymous data* to just those data elements that are relevant for the purposes.

- An **audit trail** could be created that documents the decision to access *split-off additional information*, its justification, and its responsible decision maker.
- While Recital 29 states that it is possible that the *additional information* is kept by the same controller, instituting an **independent internal entity** or an external **(trusted) third party to guard** and technically control **access to the split-off additional information**<sup>133</sup> provides an even stronger separation. These entities can then better defend the interests of data subjects, potentially even against the interests of the controller.
- Additional organizational measures can ensure that the personnel dealing with these tasks is **aware of the correct behavior** (e.g., via training) and is possibly **legally bound** (e.g., through a formal agreement to follow the above rules and procedures).

(ii) **Measures to protect the strictly pseudonymous data:**

While not explicitly stated in Art. 4(5) GDPR, controllers (and processors) shall also implement technical and organizational measures to protect the *strictly pseudonymous data*. These measures aim at preventing (direct) identification of data subjects in these *pseudonymous data*.

- The key measure to prevent (direct) identification of data subjects in the pseudonymized data is a **sufficient data pseudonymization** that is far-reaching enough to prevent direct identification. For example, a data pseudonymization that only removes unique handles from the data may be insufficient since direct

---

<sup>133</sup> Note that this does not necessarily mean that the third party actually stores the additional information. It may suffice that the third party holds a key that is necessary to decrypt the additional information. This could for example be achieved by the controller encrypting the additional information with the public key of the third party.

identification of data subject is still possible based on unique values or combinations thereof.

- Pseudonymous data are still personal data and therefore require **confidentiality**. This excludes any unauthorized external or internal party from accessing the data. Confidentiality measures typically include an **access control system** that includes authentication, authorization and maybe logging of access<sup>134</sup>.
- The controller should generally **keep the group of persons** assigned to work on the **pseudonymized data distinct from** those authorized to access the **split-off additional information**. This helps to impose restrictions on re-identification: For example, this makes it possible to restrict the amount of pseudonymous data that is being re-identified to a necessary subset; or it permits to limit re-identification to only selected data subjects. If a single person had access to both, all the *pseudonymized data* and *all the split-off additional information*, such restrictions become very difficult or impossible to implement.
- When determining the recipients to whom the *pseudonymous data* is disclosed, if necessary and possible, a controller could verify potential **motivations to re-identify** the pseudonymous data.

Where recipients are persons, a **close relationship with the data subjects** could be an indication of a potential motivation, such as curiosity. For instance, the fact that an employee is working with pseudonymous data about a group of persons to which she belongs or once belonged to, could point to a motivation of finding out who is behind certain pseudonymous data.

Similarly, where the recipient is a commercial enterprise who could **identify potential customers** in the pseudonymous data, a controller may want to verify whether a particular motivation for re-identification exists.

- Such vetting could also be used to identify personnel likely to possess **specific knowledge** about data subjects which permits to recognize (i.e., identify) persons in the data set. Again, a relationship between the personnel and data subjects could be an indicator.
- Since it is probably unfeasible to determine what knowledge personnel could possibly possess about data subjects, a controller may consider to implement ways for employees to **declare a possible “conflict of interest”** and thus avoid to work with certain data records. These can then be processed by other employees who do not have such a conflict of interest. Such a conflict of interest may for example be recognized by the fact that a data subject resides in the same general area as the employee processing the data.
- In a similar fashion, a controller can try to **assign data to work on in a way to reduce the potential of employees recognizing data subjects**. For example, a national enterprise can assign data records from one geographic region to be processed by personnel from another geographic region to render it less likely that data subjects are acquainted with personnel.

---

<sup>134</sup> Note that a logging that becomes a surveillance of personnel can also be problematic from a data protection point of view, here with the data subjects being the employees.

- The controller should consider to specify a **procedure to handle the case where an employee recognized** (i.e., identifies) **a data subject** in spite of the measures taken. The employee should report such a fact to the controller and be obliged to non-disclosure. The controller should then take steps to control possible damage arising from the identification to the data subject<sup>135</sup>. Further, it may be considered to notify the concerned data subject of this “breach”<sup>136</sup>.
- **User interfaces** used by personnel should be designed such as to show only those data elements that are necessary for the processing step at hand. By showing only a subset of a data elements, the probability of recognizing (i.e., identifying) a person is reduced. If processing steps can be completely automated without showing any data in the user interface, the possibility of recognition is eliminated altogether.
- Personnel who has access to the pseudonymous data should be made **aware** that the identification of persons in the data is not permitted. This can for example be achieved by **training** or through a **contractual obligation** with the employees.
- To separate the pseudonymous data from **additional information**<sup>137</sup> **that exists externally**, measures shall prevent that:
  - pseudonymous data can leave the (controlled) premises of the controller (e.g., by personnel taking copies home on a USB stick),
  - external data (i.e., additional information suited to identify data subjects) can be accessed on or copied to the computing systems where the pseudonymous data resides, and
  - software suitable for linking the pseudonymous data to other data sets (i.e., additional information) can be installed or used<sup>138</sup> on the computing systems where pseudonymous data reside.

---

<sup>135</sup> An obvious example is that the concerned employee stops any further access to the personal data record as soon as the identification is suspected or recognized. This may limit the amount of information learned from the identification.

<sup>136</sup> At the time of writing (January 2021), the European Data Protection Board is expected to pronounce itself on the topic of these kinds of “breaches”--at least in the context of anonymization.

<sup>137</sup> Note that this is different from the split-off additional information that is created as an output of data pseudonymization.

<sup>138</sup> Note that so called “portable” software does not require installation but can be directly used for example from a USB stick.



## 4.8 Different types of (re-) identification

There are different kinds of (re-)identifying data subjects in *pseudonymous data*. The various possibilities are illustrated in Figure 28.

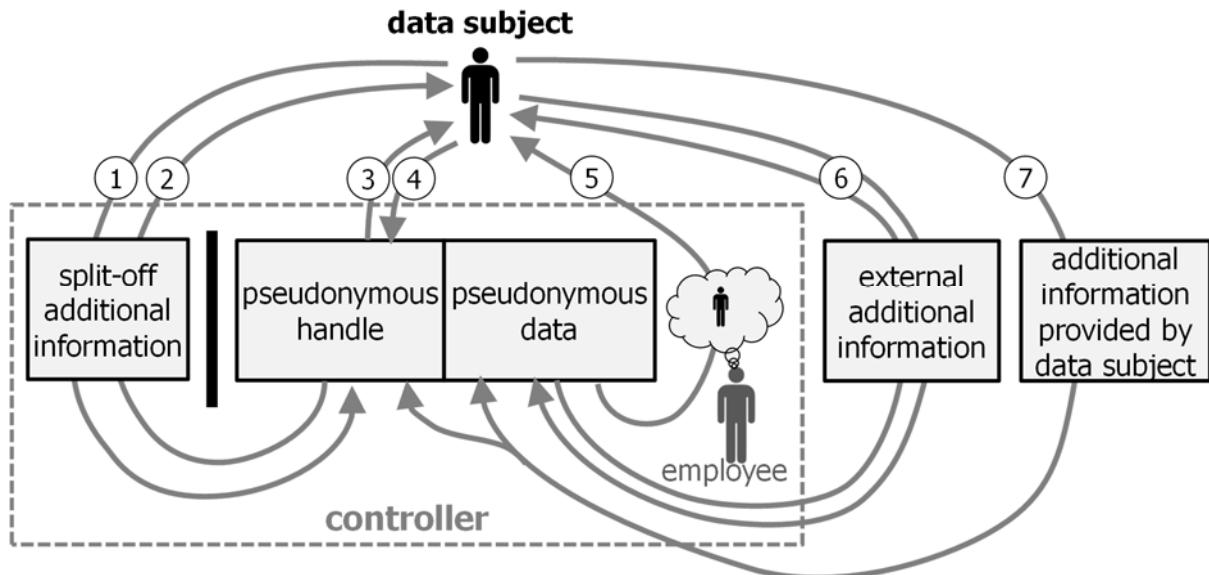


Figure 28: Different types of identification from the point of view of the controller.

The various kinds of identification are described in the following:

### (1) Locating pseudonymous data record associated with a given data subject:

In this scenario, the controller uses the **split-off additional information** in order to find the **pseudonymous handle** belonging to a known data subject. This then permits to locate the corresponding pseudonymous data record. This kind of identification can be supported with both, *bi-directional* and *one-directional additional information*.

This kind of identification is for example required for the processing of data subject right invocations such as the right of access (see Art. 15 GDPR). Here, the controller needs to be able to locate the data stored about the data subject.

### (2) Locating a data subject associated with a given pseudonymous data record:

This kind of identification inverses the direction of the previous one (1). Here, the *split-off additional information* is used to locate the identifying data elements belonging to a pseudonymous handle. This is type of re-identification is evidently only possible with *bi-directional additional information*.

This kind of identification is used when the purposes of processes require re-identification of data subjects. A controller then needs to re-identify a data subject belonging to a given record in the pseudonymous data. For example, the processing of pseudonymous health data may identify that a specific data subject suffers from a medical condition that requires rapid medical attention or intervention. The controller can then use the **split-off additional information** to obtain contact information for the concerned data subject.

### **(3) Pseudonym inversion attack:**

In some cases, flaws in the design of *pseudonymous handles* can be exploited to identify the data subject solely based on the *pseudonymous handle*. This is for example the case when a *pseudonymous handle* is created by a one-way function<sup>139</sup> (i.e., a hash or cryptographic digest) that is known<sup>140</sup> to an attacker and where the possible input values to the function are limited. Assume for example that a pseudonym is created by computing a cryptographic digest (such as *sha1*) from a 6-digit customer number. Since *sha1* is a one-way function, it may be thought that it was unfeasible to compute the customer number from the *pseudonymous handle*. Knowing the pseudonym creation function, however, and considering that there are only 1,000,000 possible customer numbers, it is possible even with very limited computing power to compute the *pseudonymous handles* for all possible customer numbers. This basically creates a lookup table that can be used to invert the pseudonym creation function. Attackers can thus exploit this flaw in the design of the pseudonym creation to identify data subjects directly from the *pseudonymous handles* themselves. Such an attack has for example been reported as having been used for the re-identification of published “anonymous” data about taxi rides<sup>141</sup>.

### **(4) Pseudonym creation attack for known data subjects:**

A similar attack that works in the opposite direction is the computation of the *pseudonymous handle* of a known data subject. Assume that an attacker knows the one-way function that was used to create the pseudonymous handle, as well as the input to this function (such as the data subject’s customer number). The attacker can then readily create the pseudonymous handle of the data subject and locate the associated record in the *pseudonymous data*. Note that this works even if there is an unlimited number of possible input values to the one-way function.

Note that the above two cases, (3) and (4), of (re-)identification are possible without the use of *additional information*. Strictly speaking, this means that the data that was called “pseudonymous data” are in fact not (strictly) pseudonymous in the legal sense. These kinds of identification should thus never occur in practice. If they are anyhow possible, it is likely due to flaws in the processing design.

### **(5) Unexpected recognition of data subject by personnel:**

In this kind of re-identification, an employee who is authorized to access the *pseudonymous data* knows the data subject and recognizes its identity based on a unique value or a unique combination of values in the *strictly pseudonymized data*.

This kind of identification is incompatible with the definition of *pseudonymization* that prohibits that *strictly pseudonymous data* supports *direct identification* of data subjects. While undesirable, it may be difficult to avoid this kind of identification in all cases. If they occur unexpectedly and in spite of the implemented technical and organizational measures, they can be handled by the controller as a kind of “breach”.

---

<sup>139</sup> According to Wikipedia, “In computer science, a one-way function is a function that is easy to compute on every input, but hard to invert given the image of a random input. Here, “easy” and “hard” are to be understood in the sense of computational complexity theory, specifically the theory of polynomial time problems.”, see [https://en.wikipedia.org/wiki/One-way\\_function](https://en.wikipedia.org/wiki/One-way_function) (last visited 28/1/2021).

<sup>140</sup> A one-way function based on a secret known only to the controller prevents that the function is known to the attacker.

<sup>141</sup> See footnote 55.

#### **(6) Indirect identification attack through the linking with external additional information:**

(6) represents two kinds of identification that operate in opposite directions. Here, externally existing *additional information* is used to indirectly identify data subjects in the *pseudonymous data*. The external *additional information* can then typically be linked to the *pseudonymous data* by exploiting unique combinations of values. The *additional information* can then either contain directly identifying data elements (such as name or e-mail address) or can again be linked to other data sets that then identify data subjects. Note that it is very difficult for a controller to predict whether external *additional information* suitable for such identification exists.

The definition of *pseudonymization* mandates to prevent any kind of *identification* and therefore prohibits the kinds of identification. Controllers typically implement confidentiality measures for the strictly pseudonymous data to prevent such identification. Training and legal obligations that prevent intended recipients to attempt such identifications are other examples of measures a controller may consider implementing.

#### **(7) Locating a pseudonymous data record based on additional information provided by the data subject:**

In this case, ***additional information provided by the data subject*** permits the controller to locate the corresponding pseudonymous data record. This case does not require the use of any *split-off additional information* stored by the controller. Instead, the *additional information* consists either of the *pseudonymous handle* or a combination of attribute values that can be uniquely matched to the pseudonymous data.

This kind of identification is foreseen in Art. 11(2) GDPR for the case where the controller can demonstrate that it is unable to identify data subjects who are invoking one of their rights.

### **4.9 Pseudonymization and Art. 11 GDPR**

While not explicitly mentioning *pseudonymization*, Art. 11 GDPR is highly relevant for it. This is evident when Art. 11

- specifies when (*split-off*) *additional information* needs to be kept by the controller (in its paragraph 1) and when
- it addresses the case where controllers can demonstrate their inability to identify data subjects (in its paragraph 2).

The latter case typically occurs when *pseudonymization* without storing *split-off additional information* is used.

This section therefore analyzes Art. 11 in more detail. The analysis is structured along the different elements of Art 11 that are made explicit in the following:

- (1) (Paragraph 1) *If 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.*
- (2) (Paragraph 2) <sup>1</sup>*Where, in cases referred to in paragraph 1 of this Article, the controller is able to demonstrate that it is not in a position to identify the data subject, ..*
- (3) *.. the controller shall inform the data subject accordingly, if possible. ..*
- (4) *.. In such cases, Articles 15 to 20 shall not apply ..*

- (5) .. except where the data subject, for the purpose of exercising his or her rights under those articles, provides additional information enabling his or her identification.

The remaining elements of Art. 11 are then used in a somewhat different order to structure the analysis as follows:

- (1) Art. 11 states that controllers who use *pseudonymization* are not obliged to store (*split-off additional information*) unless they need it for their own purposes. To understand this better, **section 4.9.1** analyses what kinds of purposes require what kinds of *split-off additional information*. On this basis, it distinguishes three types of possible *additional information* and consequently three types of *pseudonymization*.
- (2) Here, Art. 11 describes a condition where no additional information is stored and the controller is able to demonstrate that it is not able to identify data subjects. This condition is analysed in detail in **section 4.9.2** for the different types of *pseudonymization* that were distinguished in the previous section.
- (5) For the cases where condition (2) holds, Art. 11 states that **additional information** can also be **provided by data subjects**. **Section 4.9.3** discusses the requirements for *additional information* provided by data subjects to be suitable for such identification.
  - Additional information could also be provided illegitimately by imposters of the legitimate data subject. Therefore the **trustworthiness of the provided additional information** is important. This is being discussed in **section 4.9.4**.
  - In this context, controllers can preventively (typically during data collection) provide data subjects with suitable *additional information*. Data subjects can then present the received *additional information* later on when invoking data subject rights. This kind of *additional information* typically takes the form of **pseudonymous credentials**. These are discussed in **Section 4.9.5**.
- **Section 4.9.6** summarizes the cases across (2) and (5) **when identification of data subject is possible** in a way that data subject rights can be granted. On this basis, it also interprets the obligation (3) to inform data subjects.
- (4) **Section 4.9.7** discusses exactly which obligations are being waived according to Art. 11.
- Finally, **Section 4.9.8** describes which data subject rights require re-identification and which do not.

#### 4.9.1 Different types of additional information during pseudonymization

Since *strictly pseudonymous data* does not permit identification without *additional information*, the question poses itself, when a controller must store *additional information*. Art. 11(1) GDPR addresses this question:

“If 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.”

In other words, controllers need to store the *split-off additional information* only in those cases where they (still) need them for their own stated purposes of processing. This means that controllers do not need to store *additional information*, for example, for being able to process data subject right requests (see Chapter 3 GDPR) or to be able to communicate a personal data breach to the data subject (see Art. 34 GDPR).

There seem to be three cases of how a controller's specified purposes of processing require *additional information*:

- (i) **Reversibly pseudonymized data with bi-directional additional information:** Here, **re-identification** is necessary for the purposes and therefore, **bi-directional additional information** is required. This case applies for example in the usage scenarios of Figure 14 and Figure 16. Referring to Figure 28 above, the controller has access to the identification methods (1) and (2).
- (ii) **Irreversibly pseudonymized data with one-directional additional information:** Here, only **one-directional additional information** is required for the purposes of processing. This is for example the case when only new data of already known data subjects has to be integrated in a pseudonymous data set (see scenario of Figure 17). In this case, the purposes **do not require to re-identify** a data subject based on its *pseudonymous handle*. Referring to Figure 28 above, the controller thus loses access to the identification method (2) that "inverses" the data *pseudonymization*. Controllers still have access to identification method (1), however, i.e. they can locate the *pseudonymous data* belonging to a known data subject.
- (iii) **Irreversibly pseudonymized data without any additional information:** Here, the purposes of processing require **no additional information**. This is the case when **no re-identification is necessary** and no data about existing data subject is acquired at a later point in time and needs to be integrated into the existing *pseudonymous data*. This is represented in the scenario of Figure 15 above. Referring to Figure 28 above, the controller thus lacks access to both methods, (1) and (2). Compared to the previous case, even if a data subject is known (e.g., by a unique handle), the controller is now unable to autonomously locate the corresponding pseudonymous data.

Based on the different kinds of *additional information*, these three cases represent different degrees of identifiability of data subjects. Embedded in a wider context that includes also *identified* and *anonymous data*, the three cases are shown in Figure 29.

	(i)	(ii)	(iii)		
<b>Type of data</b>	<b>identified data</b>	<b>strictly pseudonymous data</b>	<b>strictly pseudonymous data</b>	<b>strictly pseudonymous data</b>	<b>anonymous data</b>
<b>Split-off additional information kept by controller</b>	N/A	<b>bi-directional additional information</b>	<b>one-directional additional information</b>	<b>none</b>	N/A
Is personal data?	yes	yes	yes	yes	no
Potential identification of pseudonymous data	direct	indirect (with <i>additional information</i> kept by the controller or external)	indirect (with <i>additional information</i> external to the controller)	indirect (with <i>additional information</i> external to the controller)	not possible (by any actor with means reasonably likely to be used now and in the future)

Figure 29: Spectrum of identification power in different kinds of data.

#### 4.9.2 Identifying data subjects with different types of split-off additional information

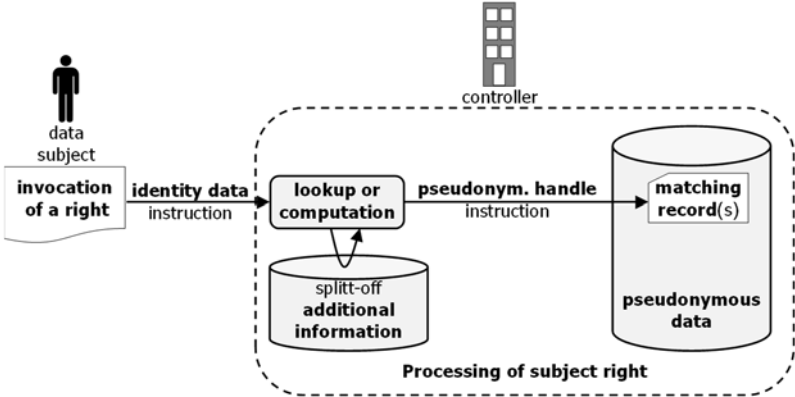
Art. 11(2) speaks of the possibility that controllers are unable to identify data subjects based on their available (*split-off*) *additional information*. This is stated in the context of processing data subject rights. The following analyses this single condition in isolation.

In more detail, Art. 11(2) GDPR uses the following wording: “Where, in cases referred to in paragraph 1 of this Article, the controller is able to demonstrate that it is **not in a position to identify the data subject**, [...]”. “Paragraph 1”, here, refers to the case where the controller refrains from storing any *additional information* since the latter is not necessary for the purposes or processing. This clarifies that the inability to identify data subjects depends on the *additional information* retained by the controller<sup>142</sup>.

In the context of the invocation of data subject rights, “identify the data subject” means to locate the record(s) in the pseudonymous data that belong to a given data subject. To identify a data subject, this must be possible based on the (i) “identity” of the data subject provided as part of the invocation and (ii) the available *split-off additional information*.

<sup>142</sup> Note that this analysis does not yet take into account the additional information that can possibly be provided by the data subject.

This process of identification is illustrated in Figure 30. Here, the data subject invokes a right (such as the right to erasure<sup>143</sup>). The invocation contains data that identify the data subject. The controller then typically uses this data to look up the matching pseudonym in the *split-off additional information*. The pseudonym is then used to locate the associated pseudonymous data record(s). This corresponds to the phrase “identify the data subject” of Art. 11(2) GDPR. Once located, the instruction (such as *erase*) that came with the invocation can be executed.



**Figure 30: Identification of data subjects during the invocation of rights.**

The figure renders it evident that the direction of identification goes from the provided identifying data elements to the pseudonym. This is possible with both, *bi-* and *one-directional additional information*.

The condition of Art. 11(2) that such identification shall (demonstrably) be impossible therefore is inapplicable to the cases of (i) *reversibly pseudonymized data* and (ii) *irreversibly pseudonymized data with one-directional additional information*. Clearly, here the *additional information* supports such identification. The condition of Art. 11(2) applies to (iii) *irreversibly pseudonymized data without any additional information*, however.

<sup>143</sup> See Art. 17 GDPR.

Figure 31 summarizes how the condition (i.e., element (2)) of Art. 11 applies in the different cases of pseudonymization.

For clarity, the table also illustrates that the condition of Art. 11(2) is different from the question whether a controller can autonomously identify data subjects in the pseudonymous data. The latter is shown in the last column of the table. It is different, because it considers “identification” in the “other direction”, i.e., from the pseudonym to some identifying data element.

		(i)	(ii)	(iii)	
<b>Type of data</b>	<b>identified data</b>	<b>strictly pseudonymous data</b>	<b>strictly pseudonymous data</b>	<b>strictly pseudonymous data</b>	<b>anonymous data</b>
<b>Split-off additional information kept by controller</b>	N/A	<b>bi-directional additional information</b>	<b>one-directional additional information</b>	<b>none</b>	N/A
Does the condition of Art. 11(2) apply?	no	no	no	yes	(yes)
Can controller identify data subject autonomously? <small>144</small>	yes	yes	no	no	no

Figure 31: Identification of data subjects during the invocation of rights.

Art. 11(2) speaks of **demonstrating** the inability to identify. In the context of pseudonymization, such demonstration has two elements:

- (i) Demonstration that no suitable *additional information* is available to the controller, and
- (ii) demonstration that the data indeed does not permit *direct identification* of data subjects (i.e., it is *strictly pseudonymous*).

Note that the former demonstration (i) concerns not only the *additional information* that was *split off* during *data pseudonymization* but includes any other data that could be used for the purpose of identification. If the *data pseudonymization* was executed correctly (ii), by definition, it is impossible to link the pseudonymous data directly with identifying data elements.

Consequently, the “demonstration” mentioned in Art. 11(2) is rather straight forward.

### 4.9.3 Additional information provided by the data subject

The above has addressed the inability to identify a data subject in the *pseudonymized data* based on the *additional information* available to the controller. Art. 11(2) foresees in these cases that suitable *additional information* can still be provided by the data subject. This possibility is discussed in more detail in this section.

---

<sup>144</sup> “Autonomously” here means without obtaining additional information from outside, e.g., from the data subject. “Identify” must here be understood to go in the other direction than the “identify” used in Art. 11(2).



As is clear from the previous section, in the case (iii) of *irreversibly pseudonymized data without any split-off additional information*, data subjects have to provide *additional information* in order to be identified. This is a pre-requisite for a controller's ability to handle data subject rights. **To be suitable**, the provided additional information must **directly and uniquely match to** a record in the ***pseudonymized data***.

Note that the objective of *data pseudonymization* is to eliminate the possibility of direct identification of data subjects. This includes the generalization of quasi-identifiers and elimination of unique values and combinations of identity-specific attributes (see *Reduction of identification potential* in section 3.7.5 above). Primarily, the according transformations prevent identification by the controller and the intended recipients, considering the information these actors have access to. Beyond this, these transformation usually also render it more difficult or impossible for data subjects to provide suitable *additional information*.

In case (iii) it is obvious that data subjects have to provide information. But also the the cases (i) and (ii), where the controller is in possession of *split-off additional information* (see Figure 30), data subjects have to provide suitable identifying data elements (i.e. *additional information*) that enable the lookup of the *pseudonymous handle* in the *split-off additional information*. This is a pre-requisite for controllers to be able to process data subject rights. Here, **to be suitable**, the provided *additional information* must **directly and uniquely match to** the identified side of ***lookup-based additional information*** or be a valid **input for formula-based additional information**.

In all these cases, the information provided by the data subject can be called *additional information*. Figure 32 summarizes the cases where data subjects can provide suitable *additional information*.

The suitable *additional information* in case (iii) requires additional analysis. In order to locate the correct *pseudonymous data* record, the controller must be able to uniquely identify it. There are two possible ways to achieve this:

- Data subjects provide **one or a combination of attribute values** that uniquely single them out in the pseudonymous data, or
- data subjects know their ***pseudonymous handle*** that matches to the pseudonym that is part of the *pseudonymized data*.

For the latter case to be feasible, the controller must obviously have communicated the *pseudonymous handle* previously to the data subject. This could for example be done during data collection from the data subject. This possibility is described in more detail in section 4.9.5 below.

		(i)	(ii)	(iii)	
<b>Type of data</b>	<b>identified data</b>	<b>strictly pseudonymous data</b>	<b>strictly pseudonymous data</b>	<b>strictly pseudonymous data</b>	<b>anonymous data</b>
<b>Split-off additional information kept by controller</b>	N/A	<b>bi-directional additional information</b>	<b>one-directional additional information</b>	<b>none</b>	N/A
Can data subject provide suitable additional information to be identified?	N/A	generally yes (typically a unique handle that matches to lookup-based additional information)	generally yes (typically a unique handle as input in the formula-based additional information)	yes, sometimes (unique combination of attributes or <i>pseudonymous credential</i> <sup>145</sup> )	no, never
Does controller need to implement data subject rights	yes	yes	yes	yes, unless no single data subject can present suitable additional information	no

Figure 32: When can data subjects provide suitable additional information?

#### 4.9.4 Trustworthiness of additional information provided by data subjects

*Additional information* could also be provided illegitimately by imposters of the legitimate data subject. Therefore the trustworthiness of the provided additional information is important; controllers need a reasonable certainty that the requestor of rights is indeed the legitimate data subject.

When imposters illegitimately invoke data subject rights by impersonating others, the legitimate data subjects are at risk of being disadvantaged or harmed. For example, a request by an imposter to access “her” data (by invoking Art. 15 GDPR) would lead to a breach of confidentiality (that is required by Art. 5(1)(f) GDPR). Similarly, an illegitimate request of erasure (by invoking Art. 17 GDPR) may deny service to the legitimate data subject and possibly lead to a loss of investment (such as the photo collection stored by the service). In the field of IT security, the ability to verify the legitimacy of a claimed identity is usually called *authentication* or *identity verification*. In the following, the former term will be used. The certainty and effort of verifying a requestor’s identity has to be proportional to the risks inherent in wrongly granting a data subject rights to an imposter.

The GDPR also states the need of authentication in (the first sentence of) its Recital 64:

“The controller should use all reasonable measures to verify the identity of a data subject who requests access, in particular in the context of online services and online identifiers.”

<sup>145</sup> *Pseudonymous credentials* are discussed below.

While the recital specifically refers to the *right of access* (see Art. 15 GDPR), it is evident that it must also apply to other data subject rights.

So the question arises how a controller can obtain reasonable certainty about the legitimate identity of a data subject who invokes a right. The options to achieve this include the following:

- (a) Evidence which links the requestor to the provided *additional information*,
- (b) the fact that the additional information provided by the data subject is unlikely to be known to anybody else, and
- (c) *additional information* that was specifically rendered verifiable<sup>146</sup>.

These are discussed in more detail in the following.

(a) The following explains the concept of such evidence by giving an example. Assume that the *additional information* provided by the data subject consists of a diagnosis and the values of some medical tests. Then, showing possession of the original documents from which such data was obtained usually provides sufficient certainty of the legitimacy. This is based on the assumption that other people may also be in possession of this information, but that it is unlikely that an imposter is in possession of the original documents.

(b) When the additional information provided by the data subject is unlikely to be known by anyone else, it is reasonable to assume that it is legitimate. This may for example apply to answers provided by a data subject to a questionnaire. In particular, free-text fields that contain the wording provided by the data subject, may be well-suited for this purpose.

(c) The additional information provided by the data subject may have been rendered verifiable by the controller or a trusted third party. This typically applies to the *pseudonymous handle* when it has been rendered verifiable in form of a credential. *Pseudonymous credentials* are discussed in the next section.

The question poses itself how stringent authentication of requesting data subjects has to be. In particular, when is it acceptable for controllers to refuse the invocation of data subject rights based on the uncertainty of authentication? The answer to this question must find a reasonable balance between data subjects' rights and the risks to the data subjects when granting the right to imposters. It is likely that controllers can only make such decisions in concrete cases and based on detailed considerations.

#### 4.9.5 Pseudonymous Credentials

In support of identifying data subjects according to Art. 11(2) GDPR, controllers can preventively provide data subjects with suitable *additional information* in the form of *pseudonymous credentials* (see definition below). This then permits identification of data subjects who present such a credential even in the case where the controller does not store *split-off additional information*<sup>147</sup>.

Note that the GDPR does not require controllers to issue such *pseudonymous credentials*<sup>148</sup>. They are interesting, however, since they illustrate that *identification* in the sense of Art. 11(2) GDPR does not

---

<sup>146</sup> Information could for example be rendered verifiable through a digital signature of a message authentication code (MAC). Such verifiable information is then comparable to a "bearer assertion" known in identity management.

<sup>147</sup> For this to work, the controller must still store the *pseudonymous handles* as part of the *strictly pseudonymous data*, however.

<sup>148</sup> This is for example explicitly stated by Marit Hansen [in German] in her commentary on Art 11 GDPR in *I. Überblick, Rn 3, Fussnote 6*, in Spiros Simitis/Gerrit Hornung/Indra Spiecker gen. Döhm (Hrsg.),

require *identified data* and can in contrast happen completely within the realm of pseudonymous processing.

*Pseudonymous credentials* are very inexpensive for controllers to produce. When data subjects invoke their rights, they significantly reduce the effort and thus cost of:

- verifying the trustworthiness of the *additional information* provided by the data subject, and
- matching it to the *pseudonymous data*.

Since controllers cannot rely on data subjects to keep their *pseudonymous credentials*, other methods to identify data subjects need to be implemented additionally, however. Where a significant volume of data subject right processing is expected, *pseudonymous credentials* may potentially lead to a significant reduction of cost.

**Definition: *Pseudonymous credential***

A *pseudonymous credential* is a piece of data that is provided by the controller to the data subject. This typically happens during data collection from the data subject (or another contact between the data subject and the controller). The piece of data has two components:

- A manner of deriving the *pseudonymous handle* of the data subject, and
- a way of verifying the legitimacy of the data and its presenter.

The following two examples shall illustrate this further.

Example 1: Here, the pseudonymous credential consists of the encryption of the pseudonymous handle with a key that is only known to the controller. The controller can then derive the pseudonymous handle through decryption and verify the legitimacy by checking that decryption results in a well-formed pseudonymous handle.

Example 2: Here, the *pseudonymous credential* consists of the *pseudonymous handle* itself together with a keyed *message authentication code* (such as an HMAC) thereof. The *pseudonymous handle* can then be directly extracted and the HMAC can be verified by computing it anew with the secret key and comparing it to the HMAC that is contained in the *pseudonymous credential*.

Beyond these examples, a wide range of other possibilities exist including the more onerous management of (one-time) passwords for each data subject by the controller (possibly connected to user accounts).

The second sentence of Recital 59 GDPR states the following in the context of data subject rights: "The controller should also provide means for requests to be made electronically, especially where personal data are processed by electronic means." Evidently, authentication is also (or particularly) necessary when requests for data subject rights are submitted electronically. Here, pseudonymous credentials linked to (pseudonymous) user accounts or issued independently may be of particular interest.

Note that *pseudonymous credentials* are not without disadvantages. Most notably, they put the burden on the data subjects to store them until the point in time when they need them.

#### 4.9.6 Summary of identifiability of data subjects and informing data subjects

Based on the detailed analysis above, the following summarizes under which conditions data subjects can be identified in a way that their data subject rights can be processed. The following list enumerates all possible cases:

- The controller stores (one- or bi-directional) *split-off additional information* and:
  - The data subject can provide a trusted identity data that matches the input side of the split-off additional information:
    - Identification is possible.
  - The data subject can provide a *pseudonymous credential*, previously issued by the controller:
    - Identification is possible.
  - The data subject can provide a trusted (combination of) value(s) that uniquely matches the *pseudonymized data*.
    - Identification is possible.
  - The data subject can provide none of the above:
    - Identification is not possible.
- The controller stores no *split-off additional information* and:
  - The data subject can provide a *pseudonymous credential*, previously issued by the controller:
    - Identification is possible.
  - The data subject can provide a trusted (combination of) value(s) that uniquely matches the *pseudonymized data*.
    - Identification is possible.
  - The data subject can provide none of the above:
    - Identification is not possible.

It is important to note that even in the case where controllers store *split-off additional information*, there is the possibility that it is technically impossible to handle data subject right invocations. More importantly, even if controllers have no *split-off additional information* at their disposition, there are anyhow cases where data subject rights can and therefore have to be processed.

The situation can be different for different data subjects and often has to be evaluated individually. Arguably, controllers need to be prepared to process data subject rights when at least one of the data subjects can successfully be identified.

Where controllers refrain from keeping *split-off additional information*, decide not to issue *pseudonymous credentials*, and significantly reduce the identification potential<sup>149</sup> in the *strictly pseudonymous data*, it can be reasonably likely that it is no longer possible to identify data subjects

---

<sup>149</sup> Reducing the identification potential of the strictly pseudonymous data reduces the probability that data subjects can provide a trusted (combination of) value(s) that uniquely matches the *pseudonymized data*.

according to Art. 11(2) and that controllers are thus relieved from implementing the processing of data subject rights.

The first sentence of Art. 11(2) GDPR states that when “the controller is able to demonstrate that it is not in a position to identify the data subject, the controller shall inform the data subject accordingly, if possible.” This was element (3) above.

In the light of the above analysis, it seems that also in the case where controllers store *split-off additional information*, data subjects need to provide suitable information in order to be identified. Also, it seems that in the case that controllers choose to issue *pseudonymous credentials*, the identification of data subjects is always possible—even if no *split-off additional information* is stored.

It may therefore be a good practice for controllers to always inform data subjects about what kind of *additional information* they need to provide to successfully invoke their rights. They may also be informed preventively under which circumstances this might fail. Since Art. 13(2)(b) and (c) GDPR already oblige controllers to inform data subjects about the existence of their rights, it seems reasonable to also inform about how to invoke them. The latter then includes the *additional information* that data subjects have to provide and in addition possibly how best to contact the controller to invoke the rights (e.g., through a specific URL for the automatic processing of rights, or through a contact point in the case of manual processing).

#### 4.9.7 Waived obligations due to inability to identify

In element (4) of the structure given above, Art. 11 GDPR acknowledges that it may be technically impossible to fulfill certain obligations of the GDPR in the case where it is demonstrably impossible to identify a data subject. In which cases this is the case has been discussed in the previous section; this section looks in more detail at how to demonstrate the inability and at exactly which obligations are being waived.

Art. 11(2) speaks of controllers demonstrating that they are not in a position to identify the data subject. As is evident from the previous section, refraining from storing *split-off additional information* is not sufficient to render it impossible to identify data subjects. Also, when data subjects can provide unique combinations of values as *additional information*, the demonstration that identification is impossible cannot be given in general but must rather be made for individual data subjects.

A technically convincing **demonstration of the inability to identify data subjects** consists of the following elements:

- If the controller has split-off additional information at disposition:
  - The fact that a data subject is unable to provide information that permits the lookup or computation of the *pseudonymous handle*, or
  - the fact that it was impossible to sufficiently establish the trustworthiness of the provided information,
- Regarding the possibility of data subjects presenting unique combinations of values:
  - The fact that the data subject was unable to provide such information,
  - the fact that it was impossible to sufficiently establish the trustworthiness of the provided information, or
  - the fact that the provided and trustworthy information did not lead to a unique match in the set of the pseudonymous data.

- If pseudonymous credentials were issued:
  - The fact that the data subject was unable to present the previously issued pseudonymous credential.

If a controller can indeed demonstrate that identification is impossible, then, according to Art. 11 GDPR, the controller is relieved from the obligations that require identification of the data subject. The article specifically states that the data subject rights of Art. 15 through 20 do not apply under this condition. But there are evidently other obligations from the GDPR that are rendered impossible by a lack of identification. Examples include withdrawal of consent<sup>150</sup> (although arguably, that is very close to the *right to erasure* of Art. 17 GDPR which is already explicitly named) and the (at least direct) communication of a personal data breach to the data subject according to Art. 34 GDPR<sup>151</sup>.

Accordingly, Art. 11(1) GDPR, instead of listing specific data subject rights, speaks more generally of “complying with this Regulation” and states clearly that controllers need not store more *additional information* solely to comply with the GDPR.

Note that Art. 11(2) GDPR also omits listing the two data subject *rights to object* (Art. 20 GDPR) and concerning *automated individual decision-making, including profiling* (Art. 21 GDPR). A detailed discussion of why these obligations were not waved was provided by Hansen<sup>152</sup>. An example she provides is the possibility of using HTTP-cookies or request headers that express the wish to opt-out and can be interpreted as exercising the *right to object*. In this case, no identification according to Art. 11(2) GDPR or verification of the trustworthiness according to Recital 64 GDPR are necessary.

Also Hornung and Wagner<sup>153</sup> discuss the obligations that are and are not waved in further detail.

#### 4.9.8 Data subject rights and the need for re-identification

Art. 11 GDPR uses the wording “identify the data subject”. It implies a linking from a known person to the associated pseudonymous data. Thus, as described in section 4.9.2 above, “identify” here assumes the meaning of “locating the associated record in the pseudonymous data”.

The term “identify” is also used in the opposite direction of linking in the GDPR. When Art. 4(1) GDPR speaks of data or information “relating to an identified or identifiable natural person”, the concepts refers to linking from data to a known person. This meaning (i.e., direction) of identify is closely related with the concepts of *identified data* and *re-identify*. It assumes the semantics of “who” and “what” coming together.

To foster a deeper understanding of data subject right processing, the following shows when *identified data* and *re-identification* are necessary and when not. It first shows prototypical cases for both and then states for all data subject rights, of which kind it is.

In the first prototypical case, *re-identification* is unnecessary. An example for this is the *right to erasure* (Art. 17 GDPR). It is illustrated in Figure 33. Here, the data subject requests that her data are being deleted [1]. The controller then “identifies” the data subject according to Art. 11 GDPR [2]. This results in the *pseudonymous handle*. The pseudonymous domain, where only “what” data can exist and “who” data is not allowed, receives only the instruction to delete the data belonging to the data

---

<sup>150</sup> Evidently in the case where consent is the legal basis for processing.

<sup>151</sup> Note that more precisely, only direct communication with the data subject about the data breach is rendered impossible, while general communication via ones web site or a newspaper advertisement is still possible.

<sup>152</sup> See footnote 145 above, RN 34.

<sup>153</sup> See footnote 101 on page 571, RN 38.

subjects *pseudonymous handle* [3]. As it should, this refrains from including any “who” data. After successful deletion, the pseudonymous domain simply signals success of the operation, without passing any data over to the domain that processes the data subject right [4]. The successful execution of the invoked right is then communicated to the data subject [5]. Evidently, there is no point in the processing where “who” and “what” data come together. In other words, the processing of the *right to erasure* does not require neither *re-identification* nor *identified data*.

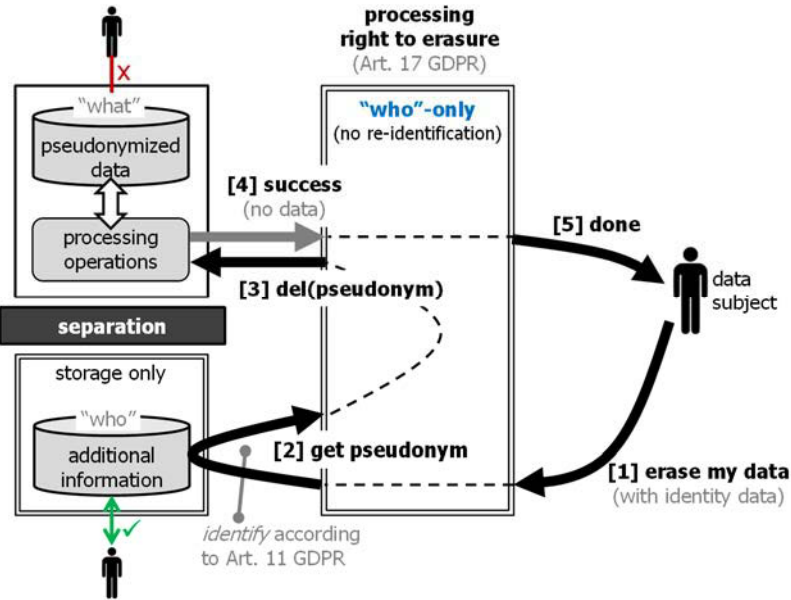


Figure 33: Processing the right to erasure does not require re-identification.



In the second prototypical case, *re-identification* is indeed necessary. An example for this is the *right of access* (Art. 15 GDPR). It is illustrated in Figure 34. Here, the data subject requests the data that are stored about her [1]. The controller then “identifies” the data subject according to Art. 11 GDPR [2]. This results in the *pseudonymous handle*. The pseudonymous domain, where only “what” data can exist and “who” data is not allowed, receives only the instruction to retrieve the data belonging to the data subjects *pseudonymous handle* [3]. As it should, this refrains from including any “who” data. After successful extraction, the pseudonymous domain passes the *pseudonymous data* belonging to the data subject to the domain that processes the invocation [4]. At this point, “who” and “what” come together, in the realm that processes the invocation. The requested data are then passed on to the data subject [5]. In this case, the processing of the *right to erasure* requires *re-identification* to take place.

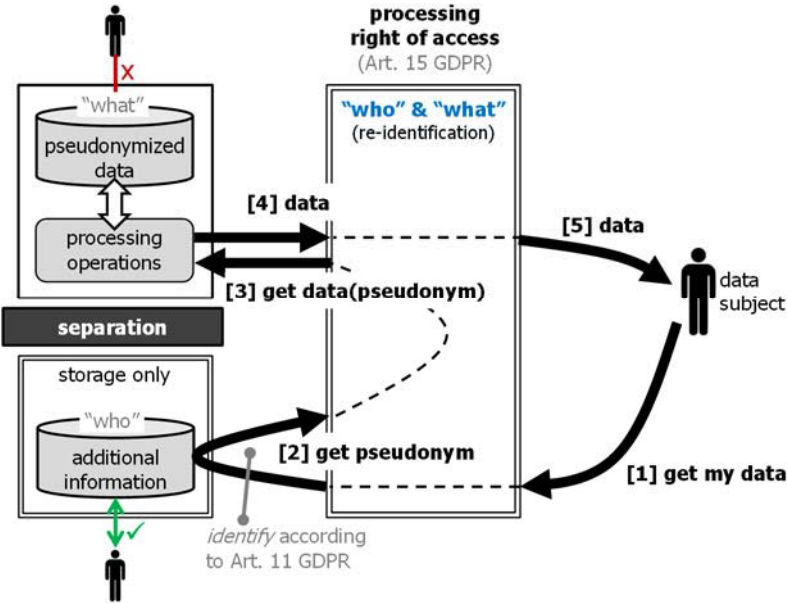


Figure 34: Processing the right of access requires re-identification.

Now that the two possible cases have been discussed in detail,

Figure 35 shows the need for using *identified data* in the domain that processes the data subject right invocation. When *identified data* is necessary, it can be either be provided by the data subject as part of the invocation or be the result of a necessary *re-identification*.

Data subject right	Need for using <i>identified data</i>
Right of access (Art. 15 GDPR)	yes (pseudonymized data must be re-identified)
Right to rectification (Art. 16 GDPR)	yes ( <i>identified data</i> is received and used as input to <i>data pseudonymization</i> )
Right to erasure (Art. 17 GDPR)	no

Right to restriction of processing (Art. 18 GDPR)	no
Right to data portability (Art. 20 GDPR)	yes (pseudonymized data must be re-identified)
Right to object (Art. 21 GDPR)	no

**Figure 35: The need for re-identification of various data subject rights.**

## 5 Anonymization

While some scholars disagree<sup>154</sup>, the GDPR considers anonymous data to be free of risk to the rights and freedoms of natural persons. Data that is inherently anonymous or results from anonymization of personal data therefore falls outside the material scope of the GDPR<sup>155 156</sup>.

Anonymous data and anonymization can be very attractive to controllers who can then avoid the cost of satisfying the obligations of the GDPR and can more freely share anonymous data in commons (e.g., in the research community) or in markets. Anonymous data can also be published, since there is no obligation to have a legal basis for the processing<sup>157</sup>, to implement protective measures such as confidentiality<sup>158</sup>, or to handle data subject right invocations.

The following discusses the concept in further detail. It first discusses how anonymization is actually defined in the GDPR. For better understanding, it then compares anonymous data with strictly pseudonymous ones. It then describes how anonymization is implemented functionally. This is followed by a discussion, whether anonymous data actually exist. To investigate this question further, some concepts relevant to the identifiability of data are defined. Finally, the section discusses options of how to deal with presumed anonymous data.

### 5.1 Definition of Anonymous

The following discusses in detail what *anonymous* actually means.

*Anonymous data* is defined in sentence 5 of Recital 26 GDPR:

“The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.”

*Anonymous data* is thus the opposite of *personal data*: Data is *anonymous* if it is not or no longer *personal*.

The properties *personal* and *anonymous* of data:  
*anonymous data* <=> *not personal data*

Recital 26 GDPR helps with the determination whether data is *personal* (and consequently also when it is *anonymous*). In particular, sentence 3 of the Recital is relevant here:

---

<sup>154</sup> An example where anonymous data pose a risk to the rights and freedoms of natural persons are automatic decisions based on anonymous profile data that discriminate against certain persons due to some bias in the anonymous data.

<sup>155</sup> See Art. 2(1) GDPR.

<sup>156</sup> See also sentence 6 of Recital 26 GDPR: “This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.”

<sup>157</sup> According to Art. 5(1)(a) and 6(1), the processing of personal data is only permissible in the presence of one of the legal bases foreseen in the GDPR.

<sup>158</sup> According to Art. 5(1)(f), confidentiality is one of the principles for the processing of personal data.

“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.”

It contains two significant elements:

- (i) **The controller or any other person** can identify the data subject, and
- (ii) account should be taken of **all the means reasonably likely to be used**.

What is meant by “means reasonably likely to be used” is further explained in sentence 4:

“To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.”

The element (i) clarifies that for data to be anonymous, it is not sufficient, if the controller is not in a position to identify data subjects. *Anonymous* requires that no other person is in a position to do so. Or in other words, if **any person** is able to identify data subjects based on the data, then the data is not *anonymous*.

This contrasts with Art. 11 GDPR which refers to situations where the controller “is able to demonstrate that it is not in a position to identify the data subject”. This was discussed in detail in the context of *pseudonymization* in section 4.9.2 above. There, the controller’s inability to identify data subjects was stated in the context of pseudonymization (with its protective technical and organizational measures) and disregarded anybody else’s ability to identify. Therefore, Art. 11 GDPR is clearly not restricted to anonymity.

*Anonymous* is thus clearly not defined relative to an actor (such as the controller). It would thus not make sense to say “it is anonymous for the controller”, since if it is not “anonymous for other actors”, it is not *anonymous* at all.

The element (ii) clarifies that the possibility to identify data subjects cannot be solely theoretical, it must be realistically be likely to happen. If an actor exists who can theoretically identify data subjects, but lacks motivation and means (such as time, knowledge, computing power, financial resources) the identification is not reasonably likely and the theoretical possibility does not render the data *personal*.

Sentence 4 of Recital 26 GDPR also adds a **temporal criterion**: “taking into consideration the available technology at the time of the processing and technological developments.” In other words, it is not sufficient for anonymity if data doesn’t allow the identification of data subjects at the time of processing, but it must also hold in the future. In other words, means reasonably likely to be used in the future must be taken into account including the following:

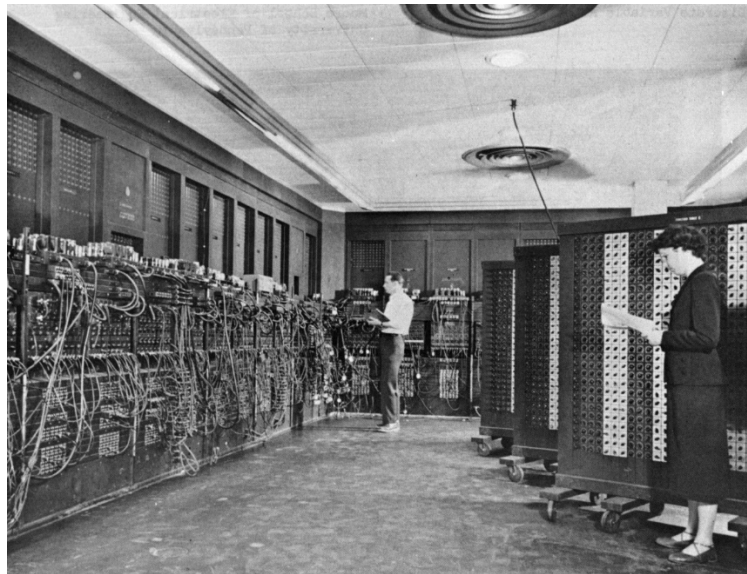
- New actors motivated in (re-)identification,
- new additional information that becomes available,
- new methodology of re-identification, and
- increased computing power (including possibly quantum computing).

If anonymous data is publicly available, this constitutes a rather high hurdle to achieve *anonymity*. The fact that the GDPR does not apply to deceased persons<sup>159</sup> suggests that the time horizon to consider may stand in relation to the life span of persons. Assuming that the youngest data subject is 20 years of age and that it is reasonably likely that a person can get 90 years old, the time horizon

---

<sup>159</sup> See Recital 27 GDPR.

is then 70 years. To foresee technological developments for this time horizon is rather challenging. It is like predicting the technical possibilities of 2020 already in 1950. This time horizon is visualized with Figure 36. Further illustrations are for example provided by William Craig<sup>160</sup>.



**Figure 36: ENIAC (Electronic Numerical Integrator And Computer) in Philadelphia, Pennsylvania. Glen Beck (background) and Betty Snyder (foreground) program the ENIAC in building 328 at the Ballistic Research Laboratory (BRL), circa 1947 to 1955, U.S. Army Photo, public domain, <https://en.wikipedia.org/wiki/File:Eniac.jpg>.**

Based on this analysis, *anonymous data* can now be defined:

**Definition: *Anonymous data***

Data is *anonymous* if any possible actor is unable to directly or indirectly (re-)identify data subjects with means reasonably likely to be used now or in the future.

Considering the definition of *identification* and Figure 5 in section 3.5 above, this means that direct or indirect linking between the anonymous data and the information assets of any actors must be impossible.

There may be one exception to this requirement of unlinkability. Namely, if the *anonymous data* were created by *anonymization* (see below) of *identified data*, actors who are in possession of the original *identified data* may be able to link with the *anonymized data*. This seems likely the case when the *anonymization* uses only truthful transformations (see section 3.7.5.3.1 above). In this case, everyone in possession of the original *identified data* is theoretically in a position to recreate the *anonymized data* and thus may be able to link<sup>161</sup>. Even if such linking is possible, however, the actor who is already in possession of the original *identified data* (or a superset thereof) would not learn any additional information from the *anonymized data*.

---

<sup>160</sup> William Craig, The History of Computers in a Nutshell, April 21, 2010, <https://www.webfx.com/blog/web-design/the-history-of-computers-in-a-nutshell/> (last visited 19/03/2021).

<sup>161</sup> One possible reason that would still prevent linking would be a random order (e.g., though shuffling) of the anonymized data records.

Note that the above definition of *anonymous*, like the definition of *anonymous* given in Recital 26 GDPR, can be seen as being a “**success state**”. This term was proposed by Mourby et al.<sup>162</sup> for the definition of *pseudonymization* in Art. 4(5) GDPR, but equally applies here to *anonymous*. Here, data is *anonymous* only if the attempts of preventing identification were successful. In other words, the state of success has been reached.

#### **Anonymization as “success state” explained**

In the GDPR, *anonymous data* are defined as data that do not or no longer permit the identification of data subjects. In other words, the state where identification is impossible must apply (in the case of originally anonymous data) or have successfully been reached (in the case of data that was created by anonymization of personal data).

This contrasts with a procedural definition that specifies a series of steps that result in *anonymous data*. For example, such steps could consist in a data acquisition procedure that avoids to collect certain identifying data elements; or it could entail the elimination of certain identifying data elements from originally personal data.

Defining *anonymous* in terms of a “success state” is very different from a procedural definition: While a procedure can always be followed successfully, the “success state” based definition fails to specify how the “success state” can actually be reached or whether it is reachable at all<sup>163</sup>.

Demonstrating that data is not *anonymous* is as easy as finding one way to identify at least one data subject. In contrast, showing that data is indeed *anonymous* is much more difficult. In particular, one needs to show that nobody possesses the know-how, additional information, and other means to identify any of the data subjects now or in the future.

This difficulty to determine whether data is indeed *anonymous*, i.e., that the “success state” where identification is not possible holds or has been reached, creates uncertainty. In many cases, it may be impossible to know for certain whether data is *anonymous* or *personal*.

The fact that a controller makes every possible effort to anonymize data according to the current state of the art is irrelevant to determine whether the data are anonymous under the GDPR. Similarly, if a controller cannot imagine any way to identify data subjects, it is not certain that the data is indeed *anonymous*. The only thing that matters in the definition of the term is whether the often elusive “success state” indeed holds or has been reached.

At first sight, one may think that the limitation to “means reasonably likely to be used” in Recital 26 may indicate that data was anonymous if only the ways of identifying data subjects are very difficult and thus less likely. This is not the case, however. Much rather, “means reasonably likely to be used” simply excludes methods of identification that are purely theoretical (e.g., since they require infinite computing power) but does not further restricts “realistically possible” ways of identification. If an identification of data actually takes place, however unlikely it is deemed, it only proves that identification was indeed “realistically possible” and the means thus reasonably likely to be used.

---

<sup>162</sup> Mourby, M, Mackey, E, Elliot, M, Gowans, H, Wallace, SE, Bell, J, Smith, H, Aidinlis, S & Kaye, J 2018, 'Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK', *Computer Law and Security Review*, vol. 34, no. 2, pp. 222-233. <https://doi.org/10.1016/j.clsr.2018.01.002> (last visited 24/03/2021).

<sup>163</sup> See section 5.5 about whether anonymous data actually exist.

Since the GDPR does not apply to *anonymous data*, the question whether data is *anonymous* or not corresponds to the question whether data subjects are entitled to protection of their rights and freedoms. It is therefore fully in line with the basic purpose of the GDPR, i.e. to protect the rights and freedoms of natural persons (see Art. 1 GDPR), to state that the entitlement to protection is always present where the processing results in a risk, no matter whether this risk was perceived or not.

## 5.2 Comparison of anonymous with strictly pseudonymous data

For a better understanding of *anonymous*, it is helpful to look at how it is different from (*strictly*) *pseudonymous*. This is done in the present section.

The following table shows the two definitions side by side. It annotates the differences.

<p>Definition: <b><i>Anonymous data</i></b></p> <p>Data is <i>anonymous</i> if <b>any possible actor</b> is unable to directly <b>or indirectly</b> (re-)identify data subjects <b>with means reasonably likely to be used</b> now <b>or in the future</b>.</p>	<p>Definition: <b><i>Strictly pseudonymous data</i></b></p> <p>Data is <i>strictly pseudonymous</i> in the <b>context of pseudonymization</b>, if, in presence of the <b>technical and organizational measures</b> of the <i>pseudonymization</i>, the <b>intended recipients</b> are unable to <b>directly</b> identify data subjects.</p>
---	---

The following differences are evident:

- While the definition of *anonymous* is **general**, *strictly pseudonymous data* is **only defined in the limited context** of *pseudonymization* with its technical and organizational measures.
- While the definition of *anonymous* refers to **arbitrary actors**, that of *strictly pseudonymous data* limits the actors to **intended recipients**.
- While the definition of *anonymous* refers to both, **direct and indirect identification**, that of *strictly pseudonymous data* limits itself to **direct identification**.
- While the definition of *anonymous* addresses the time of processing and the **future beyond**, that of *strictly pseudonymous data* limits, that of *strictly pseudonymous data* addresses only the time of processing. In other words, while *anonymous* uses an open temporal horizon, *strictly pseudonymous* uses a limited temporal horizon.

Note that the definition of *anonymous* explicitly states that only means reasonably likely to be used have to be considered. This is not explicitly stated in the definition of *strictly pseudonymous*, but it is implied by the context of pseudonymization. So there is no difference in this point.

This can be summarized by stating that both, pseudonymization and anonymization have the objective of preventing the identification of data subjects; the former does so in a controlled environment, while the latter is more ambitious by doing so in general.

Considering that

- a general property holds also in a specific context,
- intended recipients are included in “any possible actor”
- direct identification is excluded by both *anonymous* and strictly *pseudonymous*, and that
- the time horizon of *anonymous* includes that of *strictly pseudonymous*,

then it is clear that *anonymous data* are also *strictly pseudonymous* since the requirements for being *strictly pseudonymous* are a subset of the requirements for being *anonymous*.

Note that the definitions of *strictly pseudonymous* and of *anonymous* imply two different “success states” for the data. The above analysis has shown their differences in detail. In summary, the “success state” of *anonymous* is more difficult to achieve than that of *strictly pseudonymous* since it uses a superset of requirements.

The following two figures illustrate the difference between *strictly pseudonymous* and *anonymous*. First, Figure 37 shows the case of *pseudonymization* where the facilitating elements of the environment are shown in green. Then, Figure 38 shows the case of *anonymization* with the more demanding elements highlighted in red.

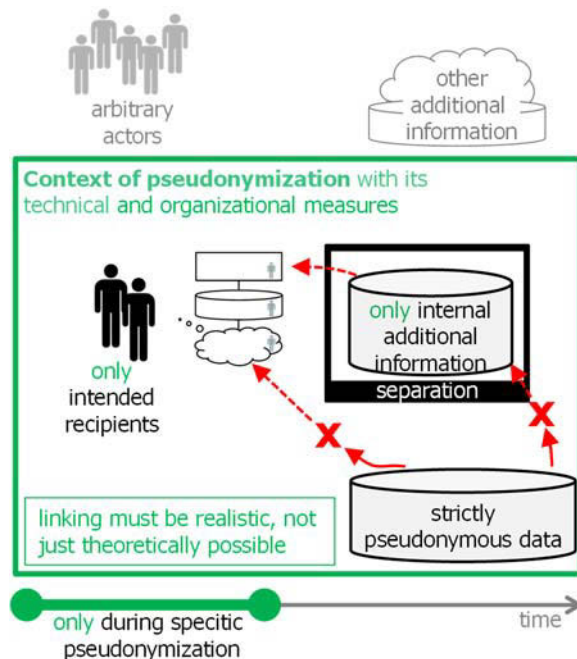


Figure 37: Data that is pseudonymous in the context of a specific pseudonymization (i.e., processing activity).

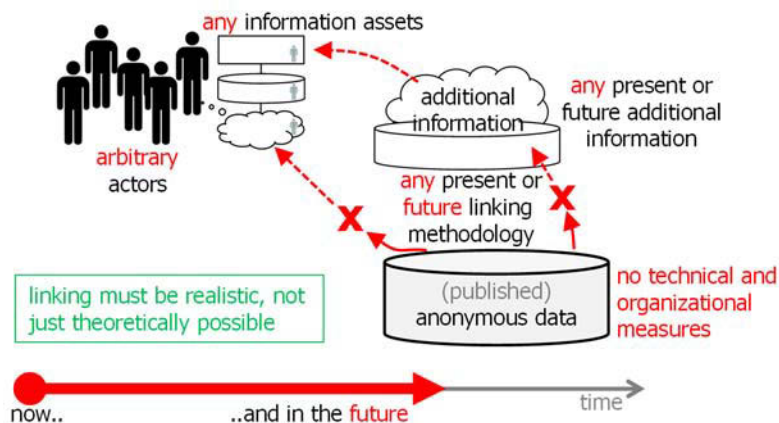


Figure 38: Anonymous data.



### 5.3 Concepts relevant to anonymization

The following defines *attempted* and *successful anonymization*, *presumed anonymous data*, as well as *successfully* and *presumably anonymized data*.

The term *anonymization techniques* is used relatively loosely in the literature in the sense that it does not guarantee that the resulting data are indeed *anonymous*. To more precisely capture the “success state” of anonymization attempts, the following definitions distinguish two concepts of “anonymization”:

Definition: (Successful) **anonymization**

*Anonymization* is a transformation that takes *personal data* as input and yields (“truly”) *anonymous data* as output. The “success state” (that identification of data subjects in the anonymous data is no longer possible) is reached.

This definition is visualized in Figure 39.

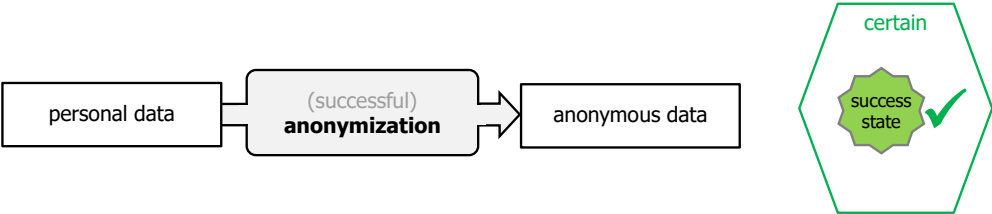


Figure 39: Anonymization.

Note that the use of the term *anonymization* thus implies the successful reaching of the necessary “success state”. Since the determination of the “success state” is often very difficult, a second concept that more closely matches actual practice is defined in the following:

Definition: **Attempted anonymization** or **anonymization attempt**

An *attempted anonymization* or an *anonymization attempt* is a transformation that takes *personal data* as input and yields *presumed anonymous data* as output. It remains unclear whether the “success state” of anonymity has indeed been reached.

This definition is visualized in

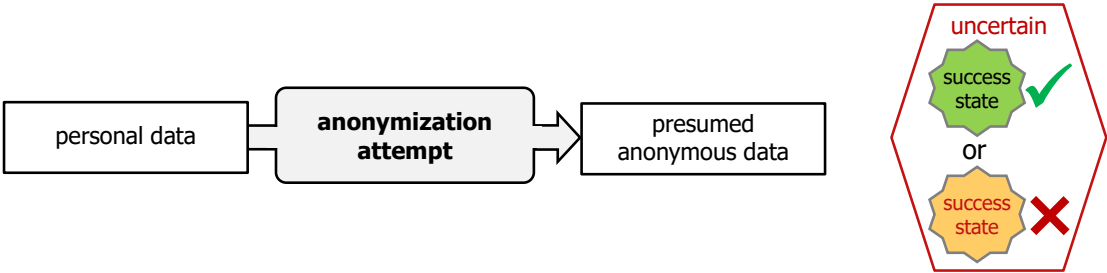


Figure 40: Attempted anonymization.

The above definition uses the term *presumed anonymous data* that is defined in the following:

Definition: **Presumed anonymous data**

*Presumed anonymized data* is data that is thought of being *anonymous* but where, due to uncertainty in the determination of the necessary “success state”, a certain risk exists that the data are actually still *personal*.

Note that to more explicitly distinguish *anonymous* from *presumed anonymous*, the term “truly” *anonymous* can be used. “Truly” *anonymous* does not add anything to *anonymous*. Much rather, it emphasizes that it is not just *presumed anonymous*.

The term *anonymized data* can be used to express that “truly” *anonymous data* has been created as the result of a successful *anonymization*:

Definition: **(Successfully) anonymized data**

*Anonymized data* is “truly” *anonymous data* that results from successful *anonymization*.

Should there be any doubt about the success of the attempted anonymization, the term *presumably anonymized data* can be used:

Definition: **Presumably anonymized data**

*Presumably anonymized data* is *presumed anonymous data* that results from an *anonymization attempt*.

## 5.4 Functional description of (successful or attempted) anonymization

This section discusses the functional implementation of *anonymization* as a subset of that of *data pseudonymization*. The functionality of *successful* and *attempted anonymization* are identical.

Functionally, anonymization is implemented by appropriate transformations which reduce the identification potential of the *personal data* (see section 3.7.5 above). The reduction is considered sufficient, when the “success state” of no longer being able to identify data subjects has been reached.

Since also *data pseudonymization* is implemented by transformations which reduce the identification potential, Figure 41 illustrates the relationship between *anonymization* and *data pseudonymization*. In particular, it shows that *anonymization* is functionally equivalent to the processing step (iv) of *data pseudonymization*. The difference lies solely in the degree of reduction of the identification potential. This was already discussed above when comparing the two “success states”.

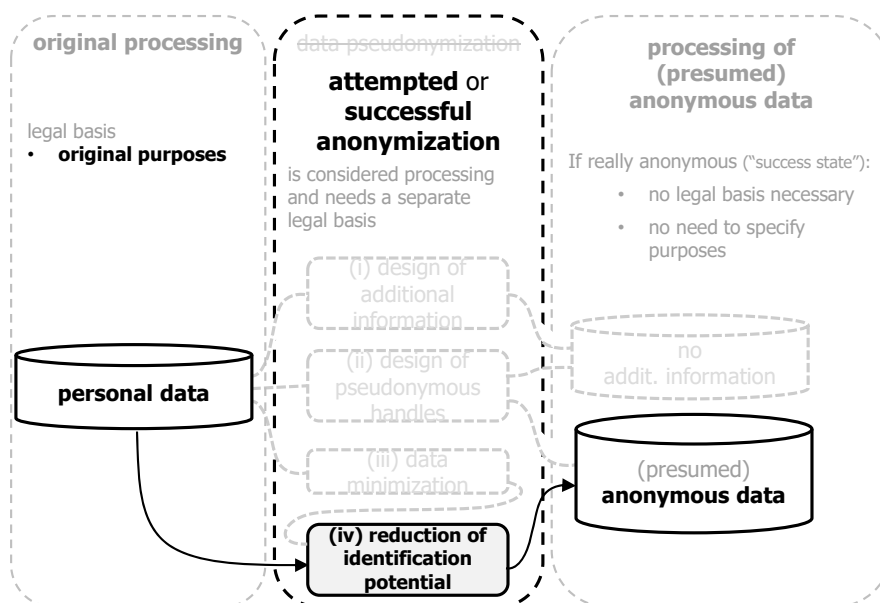


Figure 41: Anonymization as a functional subset of data pseudonymization.

It was argued earlier that the functionality of *data pseudonymization* is not sufficient to guarantee that the resulting data is *strictly pseudonymous*, i.e., that it does no longer permit the *direct identification* of data subjects. In the same way, the functionality of *attempted anonymization* does not guarantee that the “success state” of *anonymous* is actually reached.

Section 3.7.5.4 above describes how the available transformations reduce the identification potential basically gradually and that it is usually impossible to find clear indicators to determine whether the “success state” has been reached. This results in an uncertainty whether the data resulting from *attempted anonymization* are indeed *anonymous*, or, if the “success state” has not been reached, still *personal*.

## 5.5 Do anonymous data exist?

The possibility of identifying individuals in *presumed anonymous data* has received ample attention under the names of “re-identification” or “de-anonymization”. It has been widely successful and sophisticated techniques have been developed. Overviews of techniques and well-known cases are given for example by Mark Lennox<sup>164</sup>, Natasha Lomas<sup>165</sup>, Rocher et al.<sup>166</sup> and Dwork et al.<sup>167</sup>.

Some kinds of data have been found to be very difficult to anonymize. Most prominently, this holds for location data<sup>168</sup>. Here, even a generalization to country level may not be sufficient<sup>169</sup>. Also, to reduce the identification potential of data, transformation that reduce the level of detail and truthfulness of the data must be applied. The question poses itself of whether successfully anonymized data are still fit for the purposes of processing.

Many scholars have concluded that likely, anonymous data that are still useful may not exist. This was most prominently voiced by Ohm<sup>170</sup> who expresses doubt about the existence of *anonymous data* in a legal context. He states: “This mistake pervades nearly every information privacy law, regulation, and debate, yet regulators and legal scholars have paid it scant attention”. From a more technical point of view, Cynthia Dwork, the co-inventor of differential privacy, has coined the phrase “**de-identified data isn’t**” (i.e., it isn’t de-identified or it isn’t useful data)<sup>171</sup>.

---

<sup>164</sup> Mark Lennox, No such thing as anonymous data, dev.to, Oct 2, 2019, <https://dev.to/mlennox/no-such-thing-as-anonymous-data-13kk> (last visited 8/4/2021).

<sup>165</sup> Natasha Lomas, Researchers spotlight the lie of ‘anonymous’ data, TechCrunch, July 24, 2019, <https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/> (last visited 8/4/2021).

<sup>166</sup> Rocher, L., Hendrickx, J.M. & de Montjoye, YA. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* 10, 3069 (2019). <https://doi.org/10.1038/s41467-019-10933-3> (last visited 8/4/2021).

<sup>167</sup> Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, Exposed! A Survey of Attacks on Private Data, *Annual Review of Statistics and Its Application* 2017 4:1, 61-84, [https://privacytools.seas.harvard.edu/files/privacytools/files/pdf\\_02.pdf](https://privacytools.seas.harvard.edu/files/privacytools/files/pdf_02.pdf) (last visited 8/4/2021).

<sup>168</sup> See for example, de Montjoye, YA., Hidalgo, C., Verleysen, M. et al. Unique in the Crowd: The privacy bounds of human mobility. *Sci Rep* 3, 1376 (2013). <https://doi.org/10.1038/srep01376> (last visited 9/4/2021).

<sup>169</sup> Ali Farzanehfar, Florimond Houssiau, Yves-Alexandre de Montjoye, The risk of re-identification remains high even in country-scale location datasets, *Patterns*, Volume 2, Issue 3, 2021, 100204, ISSN 666-3899, <https://doi.org/10.1016/j.patter.2021.100204>.

<sup>170</sup> Ohm, Paul. (2009). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*. 57. <http://www.uclalawreview.org/pdf/57-6-3.pdf> (last visited 4/8/2021).

<sup>171</sup> Cynthia Dwork, Introduction: The Definition of Differential Privacy, Institute for Advanced Study, Four Facets of Differential Privacy, November 12, 2016, <https://youtu.be/lg-VhHlztgo?t=180> (last visited 8/4/2021).

## 5.6 Concepts relevant to the identifiability of data

The following analyses in further detail when data subjects are identifiable in data. For this purpose, it defines some relevant concepts and then distinguishes three types of data.

While according to Recital 26 GDPR, when determining whether data is personal, one has to “tak[e] into consideration the available technology at the time of the processing and technological developments”. In contrast, the concepts defined here consider the situation at a fixed point in time. This renders it possible to reason about how a situation changes over time.

**Definition: *Identification difficulty of data***

The *identification difficulty of data* captures how difficult it is to identify persons in the data set. The *identification potential of a data set* depends on the attributes it contains and on the level of detail (or generalization) and truthfulness (i.e., absence of possibly random error) of these attributes. It can be increased by applying transformations that were surveyed in section 3.7.5 above. The *identification difficulty* is the opposite of the *identification potential* of a data set.

This concept is relatively vague. In particular, it is usually impossible to measure or quantify the *identification difficulty of data*. Also, the concept is likely to be multi-dimensional such that it could not be described by a single scalar value and such that the comparison of the identification difficulty of two data sets may be difficult. In spite of these shortcomings, the concept is very useful for the discussion below.

**Definition: *General identification capability***

The *general identification capability* measures the possibility of identifying data subjects in data sets. It depends among others on the known re-identification/de-anonymization methodologies, on the availability of possibly required *additional information*, on the availability of the necessary software and computing power, as well as on the motivation and resourcefulness of potential actors. Being *general*, this ability is not pertinent to a single actor but captures the situation across all possible actors. Only re-identification/de-anonymization attempts with means reasonably likely to be used shall be considered<sup>172</sup>. The concept is similar to that of a *threat landscape* in IT security.

**Definition: *Momentary identifiability of data***

The concept of *momentarily identifiable* is used to describe a data set. It puts the *identification difficulty of data* in relation with the *general identification capability*. In particular, a data set is *momentarily identifiable* if its *identification difficulty* lies below the current *general identification capability*. The adjective of *momentarily* indicates that this comparison is made at a moment in time. This contrasts with *identifiable* (sans *momentarily*) that is used in the GDPR and considers an extended period of time<sup>173</sup>. When a data set is *momentarily identifiable*, it solely means that a realistic possibility to identify data subjects exists; not that identification has actually happened.

---

<sup>172</sup> The restriction to means reasonably likely to be used comes from sentence 4 of Recital 26 GDPR.

<sup>173</sup> Note that sentences 3 and 4 of Recital 26 GDPR contain: “To determine whether a natural person is *identifiable*, ...” and “taking into consideration the available technology at the time of the processing and technological developments”. Clearly, this use of *identifiable* is not limited to a single point in time. In the present text, the concept of *momentarily identifiable* considers only the situation at a fixed point in time, while the concept of *personal data* takes also temporal developments into account.

Note that *identifiability* was also defined for data subjects (instead of data) in section 3.5 above. There, the scope was a period of time (now and in the future), rather than a point in time.

The above concepts that describe the situation in a single point in time can now be used to describe changes over time.

Figure 42 shows an example of a typical temporal development of the *general identification capability*. The capability increased monotonically over time. The increases can be sudden, for example when a new re-identification or de-anonymization methodology is discovered; or gradual, for example when computing power gradually increases, cost of the necessary computing decreases, or the relevant available additional information gradually reaches out to more and more aspects of life.

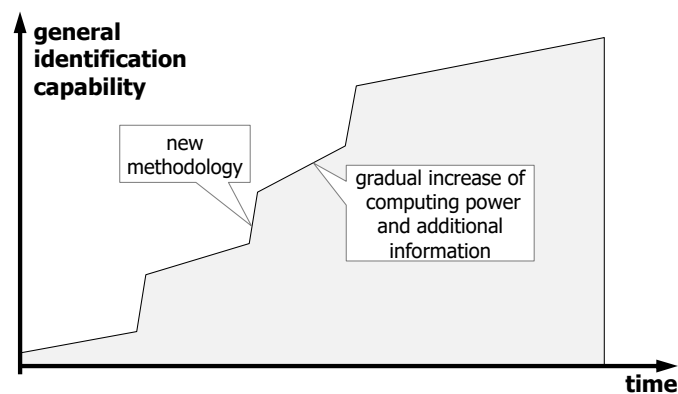


Figure 42: Example of identification capability over time.

In contrast, the *identification difficulty* of a given data set remains constant over time (unless it is modified and then results in a new, different data set).

When comparing these properties in terms of the *momentary identifiability*, the time period to consider starts with the creation of the data set. It then becomes irrelevant when the data set is destroyed and thus is no longer available to anyone.

When data is published, this may never happen, however. The definition of *anonymous* in Recital 26 GDPR does not foresee any temporal limitation of the “technological developments”. By the letter, this means that data must be considered to be personal even in the case where it will become technologically possible to identify data subjects only several generations into the future. Practically, it may be reasonable to consider a limited period of time, however. An indication may be found in Recital 27 which states that the GDPR does not apply to deceased persons<sup>174</sup>. Where the data does not provide information about persons related to the data subjects<sup>175</sup>, a generation may thus be the time to consider.

The following looks at the **possible cases of momentary identifiability** of data over the relevant period of time. The figures list only the destruction of data; in the case of publication, a further time horizon would have to be considered.

---

<sup>174</sup> Note however, that the same Recital 27 GDPR states that “Member States may provide for rules regarding the processing of personal data of deceased persons”.

<sup>175</sup> Note for example, that genetic data may constitute personal information also about decedents of the original data subjects.

(i) The first case is that of **continuously identifiable data** that is shown in Figure 43. Here, starting with the time of creation of the data set, the latter is continuously *momentarily identifiable*. Evidently, the data set therefore constitutes *personal data*.

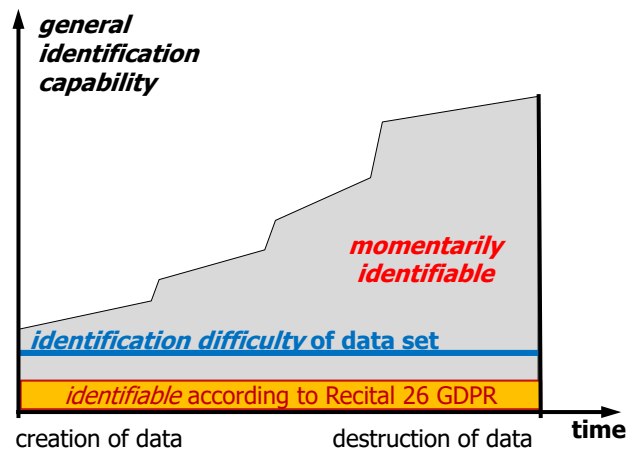


Figure 43: Continuously identifiable data.

Definition: **Continuously identifiable data**

*Continuously identifiable data* are *personal data* that are continuously *momentarily identifiable* from their creation to their destruction.

(ii) A second case is that of **eventually identifiable data** that is shown in Figure 44. Here, at the time of creation, the data set is not *momentarily identifiable*. *Momentary identifiability* occurs only starting from a later point in time that lies before the destruction of the data set.

Based on sentence 4 of Recital 26 GDPR, *eventually identifiable data* constitutes *personal data*. **It is important to understand that they are always *personal data*.** Thus, it would be incorrect to think of the data as “truly” *anonymous* until the moment when they become *momentarily identifiable*, and from then onward as *personal data*.

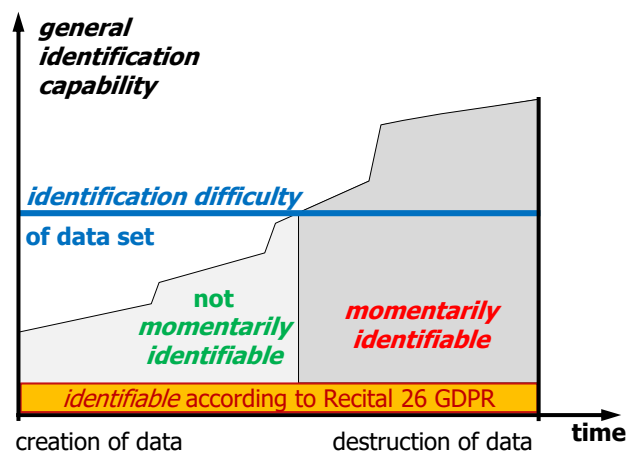


Figure 44: Eventually identifiable data.

Definition: **Eventually identifiable data**

*Eventually identifiable data* are data that are not *momentarily identifiable* at the time of their creation but become *momentarily identifiable* before their destruction.

(iii) The third case is that of **anonymous data**. It is illustrated in Figure 45. Here, the data are never *momentarily identifiable* in the period from their creation to their destruction.

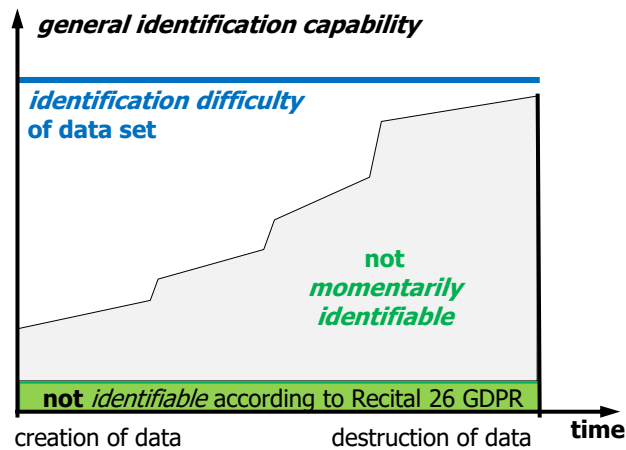


Figure 45: Anonymous data.

*Anonymous data* was already defined above in section 5.1. The present discussion and alternative definition is completely compatible with that definition. It solely focuses on the temporal aspects and uses the concept of *momentarily identifiable* that has since been defined.

Alternative definition: **Anonymous data**

*Anonymous data* are data that are not *momentarily identifiable* over the complete period between their creation and their destruction.

## 5.7 Options to deal with *presumed anonymous data*?

The present section discusses how controllers can deal with the uncertainty of assessing the “success state” in terms of which (*truly*) *anonymous* is defined. It first briefly reflects on the sources of the uncertainty and then discusses the options that stand at the disposition of controllers.

*Anonymous* has been defined as a “success state” that no actor can identify data subjects in the data with means reasonably likely to be used. This same “success state” is inherent in the concept of *momentarily identifiable*. Whether the “success state” applies often depends on the possible external actors, their know-how about re-identification/de-anonymization methods, the additional information they have at their disposition, the resources they are likely to employ, and the state of technology potentially decades into the future. It is likely impossible for controllers to obtain sufficient information about these factors.

While there are technical indicators to measure certain aspects of *anonymity*, none of them is general enough to provide guarantees to reach the “success state”. For example, *K-anonymity* is an indicator whether certain types of re-identification/de-anonymization are possible. But it has been found that the indicator is not sufficient to guarantee the “success state” and consequently, additional complimentary indicators such as *L-diversity* have been proposed. Even the combination of all these indicators fails to provide guarantees. For example, these indicators are often used only for *quasi-*

*identifiers*, leaving the possibility of identifying data subjects based on unique combinations of *identity-relevant properties*.

The best indicator (or better theory) on how to reach the “success state” is arguably *differential privacy*. This is due to the fact that it puts all data into scope (not just a subset thereof) and makes no assumptions about the information or capabilities available to external actors. But it has also been shown that *differential privacy* is very difficult to apply in practise (section 3.7.5.3.2.3 above).

Consequently, the evaluation of “success states” is in many cases a highly difficult task for controllers and the resulting assessment is often plagued by a significant level of uncertainty. The following looks in more detail about how controllers can best manage this uncertainty and the resulting risks.

Controllers must decide even before the time of creation of a data set (through data collection from data subjects or by derivation from another data set) what kind of data they are dealing with. Even if the controller presumes that the data are *anonymous*, due to the uncertainty, they could actually be one of the following:

- *continuously identifiable*,
- *eventually identifiable*, or
- “truly” *anonymous*.

In the former two cases, the data are *personal* and the GDPR is applicable; in the latter case it isn't. Considering the potentially significant uncertainty in the assessment of the type of data, the following two risks emerge:

- Controllers erroneously classify *personal data* as *anonymous* and consequently fail to comply with the requirements of the GDPR, and
- controllers, possibly out of prudence, treat *anonymous data* as if they were personal and make an unnecessary effort of implementing the requirements of the GDPR.

Figure 46 gives an overview of all possible cases. The lines represent the possible actual data types; the columns show the decision by the controllers whether to treat the as *anonymous* or *personal data*, respectively.

Every cell shows the consequences of the controller's decision. It includes the following aspects:

- Whether the controller's decision represents the correct classification of the data;
- whether the controller's treatment is compliant with the GDPR;
- whether there is a potential of irreparable damage for data subjects or at least for an irreparable infringement of their rights and freedoms;
- the obligations that a controller faces (in most cases at the point of time when it becomes clear that the data is indeed personal).



<b>Data:</b>	<b>Treat as anonymous data</b>	<b>Treat as personal data</b>
<b>“truly” anonymous</b>	<p>[correct classification]</p> <p><b>GDPR-compliant</b></p> <p><b>Obligations according to GDPR:</b> in some cases, a Data Protection Impact Assessment (DPIA) is required before anonymization<sup>176</sup></p>	<p>[incorrect classification]</p> <p><b>GDPR-compliant</b> (extra effort is allowed)</p> <p><b>Obligations according to GDPR:</b> none, but implementation of measures insures against consequences of classification error.</p>
<b>eventually identifiable</b>	<p>[incorrect classification]</p> <p><b>GDPR violation</b></p> <p><b>Potentially irreparable damage for data subjects</b></p> <p><b>Obligations according to GDPR:</b> Mandatory damage control, possible termination of processing, consequences of GDPR violation and potential liability claims</p>	<p>[correct classification]</p> <p><b>GDPR-compliant</b></p> <p><b>Obligations according to GDPR:</b> Implementation of technical and organizational measures.</p>
<b>continuously identifiable</b>	<p>[incorrect classification]</p> <p><b>GDPR violation</b></p> <p><b>Potentially irreparable damage for data subjects</b></p> <p><b>Obligations according to GDPR:</b> Mandatory damage control, possible termination of processing, consequences of GDPR violation and potential liability claims</p>	<p>[correct classification]</p> <p><b>GDPR-compliant</b></p> <p><b>Obligations according to GDPR:</b> Implementation of technical and organizational measures.</p>

Figure 46: The different options available to controllers to deal with *presumed anonymous data*.

<sup>176</sup> For example, in Germany, in the private sector, the list according to Art. 35(4) GDPR of processing applications that require a Data Protection Impact Assessment, include the anonymization of special categories (according to Art. 9 GDPR) of personal data. See Nr. 15, page 4, [https://www.bfdi.bund.de/SharedDocs/Downloads/DE/Datenschutz/Liste\\_VerarbeitungsvorgaengeDSK.pdf?blob=publicationFile&v=4](https://www.bfdi.bund.de/SharedDocs/Downloads/DE/Datenschutz/Liste_VerarbeitungsvorgaengeDSK.pdf?blob=publicationFile&v=4) (last visited 6/5/2021).

The sequel describes in more detail the obligations facing a controller when it is discovered that the classification of data as *anonymous* was incorrect. It covers in particular the following:

- (1) What are examples for the potentially irreparable damage and disadvantages for data subjects?
- (2) What are the possible consequences of a GDPR violation?
- (3) In what does the mandatory damage control<sup>177</sup> consist?
- (4) How substantial is the effort of treating presumed anonymous data as being personal when there is any doubt?

### **5.7.1 Potential damage and disadvantage to data subjects**

The very objective of the GDPR is to protect the rights and freedoms of data subjects when their personal data is being processed by controllers. When personal data is processed without observing the obligations of the GDPR, data subjects are therefore deprived of their rights and freedoms.

For example, when data is erroneously presumed to be *anonymous*, data subjects are typically not informed about the processing of their data (lack of transparency), and thus cannot exercise their rights, such as objecting to the processing on the basis of their specific situation. Beyond this, the data may not be managed with the safeguards prescribed by the GDPR. This deprives data subjects of the necessary protection and exposes them to increased risks of disadvantage or damage. Further, when controllers fail to have a legitimate legal basis, the power imbalance between controller and data subject is tilted all the way in favour of the controller.

It is evident that the above consequences cannot be remedied in retrospect.

Beyond the above impact on the rights and freedoms of data subjects, data subjects can experience irreparable damage. Assume for example that unsuccessfully anonymized medical data about some sensitive disease (such as HIV) get published and later, it is found out that some of the data subjects can be identified. In consequence, these data subjects may suffer highly adverse consequences at their workplace, in their career, as well as their relationships.

It is also here evident that such damage once done is irreversible and beyond remediation.

### **5.7.2 Consequences of a GDPR violation**

In the options above, the GDPR was violated when *personal data* was treated as if it were *anonymous*. In this case, the controller typically assumed that the processing was not subject to the requirements of the GDPR and did not satisfy its requirements.

This bears a risk of, e.g., administrative fines imposed by the supervisory authorities (Art. 83 GDPR), or legal compensation proceedings (Art. 82 GDPR). Also, supervisory authorities may issue administrative orders to bring processing operations into compliance with the GDPR (Art. 58(2)(d) GDPR) or they even may impose a temporary or definitive limitation including a ban on processing (Art. 58(2)(f) GDPR).

---

<sup>177</sup> *Damage control* is used in a wider sense, comprising any action that is necessary as a consequence of processing *personal data* as if they were *anonymous*. It is thus not limited to the potential damage suffered by data subjects.

For all legal action taken, the circumstances of the individual case have to be considered. This is illustrated in Art. 83(2) GDPR that lists several criteria that have to be taken into account when deciding on an administrative fine. Those most relevant to the situation at hand are reported here:

- “the intentional or negligent character of the infringement” (Art. 83(2)(b) GDPR),
- “the manner in which the infringement became known to the supervisory authority, in particular whether, and if so to what extent, the controller or processor notified the infringement” (Art. 83(2)(h) GDPR),
- “the degree of cooperation with the supervisory authority, in order to remedy the infringement and mitigate the possible adverse effects of the infringement” (Art. 83(2)(f) GDPR),
- “any action taken by the controller or processor to mitigate the damage suffered by data subjects” (Art. 83(2)(c) GDPR),

Finding out that *presumed anonymous* data are after all identifiable may happen in good faith and with neither intention nor negligence. But once controllers realize the violation, they should act responsibly and swiftly to control the damage.

Controllers need to decide whether they notify the competent supervisory authority about the GDPR violation. As stated in Art. 83(2)(h) GDPR, this may be a factor that is considered favourably by supervisory authorities. As a matter of fact, processing *personal data* as if it was *anonymous* could be considered to be a **personal data breach** according to Art. 4(12) GDPR. This is particularly the case when the personal data was accessed by unauthorized persons; and in absence of a valid legal basis (according to Art. 6 GDPR), the processing is unlawful and therefore nobody can be authorized.

According to Art. 33(1) GDPR, “[i]n the case of a *personal data breach*, the controller shall without undue delay and, where feasible, not later than 72 hours after having become aware of it, notify the personal data breach to the supervisory authority competent in accordance with Article 55, unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons.” The decision not to notify a *personal data breach* can thus only be made on the basis of a risk assessment.

Evidently, in any case, controllers have to take rapid actions to satisfy the GDPR requirements (which wouldn’t have been necessary for *anonymous* data). This is discussed in the following subsection.

### **5.7.3 Mandatory damage control when presumed anonymous data is discovered to be personal**

The following looks in further detail what obligations of the GDPR were disregarded when data was wrongly assumed to be *anonymous* and what damage control is required. For this purpose, the obligations stated in the GDPR in the chapters “Principles”, and (obligations of) “Controller and Processor” are systematically discussed.

#### *Principles:*

- *Lawfulness*  
Art. 5(1)(a) GDPR requires that the processing of personal data be lawful. To be lawful, controllers must select a legal basis that is foreseen in Art. 6(1) and 9(1) GDPR. Typically, controllers who presumed that the data was (truly) *anonymous* fail to select such a legal basis.

To handle this omission, controllers need to select a valid legal basis as soon as possible, desirably also in retrospect for past processing. A legal basis is a prerequisite for continued processing activities; a legal basis in retrospect provides a certain justification for keeping the results of already completed processing.

A prerequisite for this is to explicitly declare the purposes of processing. A legal basis must then be found for every single purpose. Should it prove impossible to find a legal basis for some or all purposes, the corresponding processing has to be terminated immediately.

The retarded selection typically restricts the selection of possible legal bases. In particular, the popular legal basis of “consent” (see Art. 6(1)(a) and 9(1)(a) GDPR) is no longer available since the possibility to communicate with data subjects typically does not exist<sup>178</sup>. In the case that a controller selects “public interest” (see Art 6(1)(e)) or “legitimate interests by the controller or a third party”, controllers have to successfully conduct and document a balancing test<sup>179</sup> that shows that the public or legitimate interests prevail over the interests, rights and freedoms of data subjects.

Where a legal basis for past and continuing processing can be found for the processing, the processing can continue and past results can be kept with a certain justification. Where this is not possible, any further processing has to be ceased and the data and results of unlawful past processing has to be deleted. Some kinds of processing cannot be “reversed”, however. For example, if in retrospect, no legal basis can be found for the disclosure of personal data to third party recipients, it is usually not possible to “undisclose” these data. In such cases, it may be advisable to seek support from the competent supervisory authority.

- *Transparency*

When controllers presume data to be *anonymous*, they typically do not inform data subjects about the processing as would be required by Art. 12 through 14 GDPR, nor do they include the processing in the *records of processing activities* required by Art. 30 GDPR.

An incomplete remediation of this situation would include to inform data subjects about the past, present, and future processing of their data. To contact data subjects may often not be possible. For example, this is the case when a controller is not in a position to identify the data subjects or obtain suitable addresses for communications.

To avoid “secretive” processing of personal data that completely evades scrutiny, information about the processing can be published, however. This can for example be done on the public web site of the controller and it can include at least the information elements required by Articles 13 and 14 GDPR including informing data subjects about their rights and providing contact information for invoking such rights.

A lack of transparency is particularly critical when a legal basis of “legitimate interests” (Art. 6(1)(f) GDPR) was chosen for at least parts of the processing. In this case, the fact that data subjects are unaware of the processing prevents them from exercising their right to object (Art. 21 GDPR) to the processing based on their specific situation that was not taken into consideration during the balancing test. In this case, controllers should make a particular effort to provide transparency and support data subject rights.

---

<sup>178</sup> This is particularly true if the controller only knows that re-identification/de-anonymization is possible but is not itself in a position to do so for all data subjects.

<sup>179</sup> See Article 29 Data Protection Working Party, Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC, WP217, Adopted on 9 April 2014.

- *Purpose limitation*  
Controllers who erroneously presume data to be *anonymous* typically disregard the principles of purpose limitation. To address this, any future processing has to be limited to those purposes for which a valid legal basis was found. Where data was processed in the past for purposes without a legal basis, any results and outputs of such processing shall be immediately erased and undone. Where this is not possible, controllers may be well advised to contact the competent supervisory authority for support.
- *Data minimization*  
Controllers who erroneously presume data to be *anonymous* typically disregard the principles of data minimization and keep all available data. To address this, all data needs to be deleted when it is not or no longer necessary for reaching the purposes for which a legal basis has been found. This may eliminate a part of data that was kept solely because it may be of interest sometime in the future.
- *Accuracy*  
Controllers are responsible to keep personal data accurate and up to date. Where data was presumed to be *anonymous* and where controllers likely cannot identify data subjects, to verify the accuracy may often not be possible. But with a very weak identifying potential of the data, the risk arising from inaccuracy to data subjects is likely also very marginal.
- *Storage limitation*  
Considering that the data was previously presumed to be anonymous, it is likely that the data is already in a form that keeps the potential of identification of data subjects to a minimum. If the identification potential can be reduced further (while still achieving the purposes for which a valid legal basis has been found), controllers are obliged to do so. A further reduction of the identification potential of the data may even swart the re-identification/de-anonymization approaches that caused the data to be recognized to be personal much rather than anonymous.
- *Integrity and confidentiality*  
While integrity, similar to accuracy, seems mostly unproblematic, confidentiality is highly critical to the damage control. When controllers presume data to be *anonymous*, they typically do not implement any technical and organizational measures to keep the data confidential. It may even be that the *presumed anonymous data* has been published or otherwise been made widely available.

In contrast to *anonymous data*, for *personal data*, controllers need to prevent unauthorized and unlawful processing (Art. 5(1)(f) GDPR). This entails the prevention of disclosure<sup>180</sup> to unauthorized recipients whose activity fails to contribute to reaching<sup>181</sup> the legitimate purposes<sup>182</sup> and of the processing.

It is typically very difficult to find a remediation for confidentiality once it is lost. One cannot “put the genie back into the bottle” or as Will Rogers put it<sup>183</sup>: “Letting the cat outta the bag is a whole lot easier than putting it back”. Particularly the publishing of data is often irreversible.

---

<sup>180</sup> According to Art. 4(2), disclosure constitutes processing.

<sup>181</sup> The principle of purpose limitation restricts the processing (including disclosure) to what is necessary to achieve the legitimate purposes.

<sup>182</sup> I.e., the purposes for which a valid legal basis has been found.

<sup>183</sup> <https://www.coolnsmart.com/quote-letting-the-cat-outta-the-bag-is-1214/> (last visited 27/4/2021).

Where data has been disclosed to third countries or international organizations, also the obligations of Chapter 5 of the GDPR have to be taken into account when determining the irreversible violation and the possible damage control effort. The fact that the third country recipients of the information may not be bound to provide safeguards for data subjects that are comparable to the GDPR may increase the damage and can impede the efforts to control it.

While past disclosure is typically not reversible<sup>184</sup>, controllers obviously must refrain from further disclosure of the data to unauthorized parties. This is usually manageable where data is used internally or by processors since here, controllers are mandated to tightly control the processing activity and the recipients of the data (see Art. 28 and 29 GDPR).

A more difficult question is how to handle the situation from now into the future where data were disclosed (in violation of the GDPR) to third party recipients. While the GDPR fails to address this question, Art. 17(2) may provide some indication in this respect. In the context of the right to erasure, it states that “[w]here the controller has made the personal data public and is obliged [...] to erase the personal data, the controller, taking account of available technology and the cost of implementation, shall take reasonable steps, including technical measures, to inform controllers which are processing the personal data that the data subject has requested the erasure by such controllers of any links to, or copy or replication of, those personal data.”

In analogy, controllers who disclosed data as “*anonymous*” to third parties should take reasonable steps to inform these recipients that the data was in fact *personal* and that also they are obliged to take damage control action. Considering that Art. 17(2) GDPR describes a situation in absence of any violation of the GDPR, it may be argued that, in presence of a violation, the controller should make an even bigger effort. Note that also the recipients who further disclosed the *presumed anonymous data* to additional recipients need to recursively propagate the necessary damage control.

#### *Obligation of controllers and processors:*

The following looks at how to control damage in respect to some specific obligations of controllers that are described in Chapter 4 of the GDPR.

- *Data protection by design and by default (Art. 25 GDPR)*  
When determining the means of processing for personal data, controllers shall already during the design phase consider the risks of the processing to the rights and freedoms of the data subjects and how these risks can be mitigated with suitable technical and organizational measures. When data was wrongly assumed to be anonymous, the processing was typically designed in disregard of this obligation. A damage control effort therefore has to identify the risks inherent in the processing and potentially re-design the processing activity in a way that provides adequate safeguards for data subjects. Such a re-design could come with a substantial disruption and cost.
- *Vetting of processors (Art. 28(1) GDPR )*  
According to Art. 28(1) GDPR, a “controller shall use only processors providing sufficient guarantees to implement appropriate technical and organisational measures [...]”. Where a controller erroneously assumed that the data was *anonymous*, such a vetting of possible

---

<sup>184</sup> It is for example not possible to make a human recipient forget disclosed facts.

processors was typically not performed. To partially remedy this omission, processing activities at unsuitable processors need to be terminated and moved to suitable processors instead. This may entail the premature and unforeseen termination of a contract with a possible financial loss on part of the controller.

- *Contractual agreements with processors (Art. 28(3) GDPR)*  
Where a processing activity makes use of processors, controllers who erroneously assumed that the data were *anonymous* have likely not included all the required clauses into the contracts with the processors. To partially remedy this situation, the contracts have to be changed accordingly. This may well entail unforeseen changes of running contracts that require re-negotiations and possibly changed remuneration.
- *Data Protection Impact Assessment (Art. 35 GDPR)*  
Where the processing entails a high risk, controllers have to conduct a Data Protection Impact Assessment (DPIA). In the cases where no DPIA is required, it is good practice for controllers to have documented the reasoning that led to the conclusion that the processing does not entail a high risk. Note that it may well be that the processing of anonymized data entails a low risk while the anonymization itself is considered a high risk. This is for example the case in Germany, where the anonymization of special categories of data (according to Art. 9 GDPR) is included by the competent supervisory authorities in the list (according to Art. 35(4) GDPR) of processing operations that require a DPIA<sup>185</sup>.
- *Designation of a Data Protection Officer (Art. 37 GDPR)*  
Controllers who incorrectly presumed data to be anonymous may not have designated a data protection officer and may now have to do so.
- *Designation of a representative for controllers not established in the Union (Art. 27 GDPR)*  
Controllers not established in the Union who incorrectly presumed data to be anonymous may not have designated a representative in the Union and may now have to do so.
- *Security of Processing (Art. 32 GDPR)*  
Controllers who incorrectly presumed data to be anonymous may not have implemented sufficient technical and organizational measures that guarantee security and may now have to do so.
- *Records of processing (Art. 30 GDPR)*  
Controllers who incorrectly presumed data to be anonymous may not have included the processing into the records of processing and may now have to do so.

#### *Summary of damage control action:*

Based on the above discussion of damage control action, the following summarizes the kinds of actions that are required. It looks at past and present processing operations:

#### *Past processing operations:*

- Create retrospective compliance (e.g., retrospectively finding a legal basis).
- Implement retarded compliance (e.g., informing data subjects about the processing, processing of data subject right invocations).
- Reverse effects of unlawful processing (e.g., deleting data and results).

---

<sup>185</sup> See footnote 176.

- Report irreversible effects of unlawful processing to the competent supervisory authority.
- Inform possible third party recipients of the need for equivalent damage control action.

*Present processing operations:*

- Stop processing until indispensable pre-requisite obligations are fulfilled (e.g., legal basis, DPIA).
- Satisfy obligations as quickly as possible during processing (e.g., designate a DPO, create more efficient processes to handle data subject rights, implement additional and improved technical and organizational measures).

The most critical aspect of the damage control action is how to handle irreversible effects of unlawful processing. This includes (but is not necessarily limited to):

- Unlawful transfer of data to third party recipients (possibly even in third countries),
- unlawful publication of data, and
- irreversible effects of unlawful processing on data subjects (such as decision making affecting data subjects<sup>186</sup>)

#### **5.7.4 Implementing GDPR requirements for presumed anonymous data**

The previous two subsections have discussed the consequences when a controller falsely treats data as *anonymous* but finds out at a later point that it is *personal* after all. This subsection looks at what exactly has to be done to “play it safe” and treat *presumed anonymous* data as *personal data*. In other words, it looks at how much an “insurance” actually costs.

There is a base effort for a controller to process personal data. It includes the following:

- Obtain a basic knowledge about the obligations of controllers,
- possibly the designation of a data protection officer (DPO),
- setting up internal procedures to handle the processing of personal data (e.g., internal approval of processing activities such that controllers can assume their responsibilities), and
- setting up a certain infrastructure (such as the *records of processing activities* according to Art. 30 GDPR).

For organizations that already process *personal data* for other purposes, this comes at a zero or marginal additional effort. The following thus focuses on specific effort that is necessary when treating *presumed anonymous data* as *personal data*. Not surprisingly, the structure of the discussion closely reflects that used for damage control efforts in the previous section.

- *Lawfulness*

For the processing of personal data to be lawful, a controller to explicitly state the purposes of processing and find a legal basis according to Art. 6(1) and 9(1) for each of them. Where controllers (attemptedly) *anonymize* personal data, the GDPR considers this also as a processing operation that requires a legal basis. In contrast to doing this as a damage control effort, where data is collected from data subjects and later (attemptedly) anonymized, consent (according to Art. 6(1)(a) and 9(1)(a)) remains a possible legal basis. For controllers familiar with the task, defining purposes and selecting legal bases are very limited efforts.

---

<sup>186</sup> An example for such decision making would be the refusal of a credit or service, or the denial of a right.



- *Transparency*  
In the case where controllers collect data directly from data subjects or process data the still permits to contact data subjects, informing data subjects according to Art. 13 or 14 GDPR, respectively, is an easy task. Where this is not the case, transparency towards data subjects has to be created in the same way as during a damage control effort: For example, a web page of the controller can inform about the processing and contain the elements requested by of Art. 14 GDPR. The processing activity also has to be added to the *records of processing activities* (according to Art. 30 GDPR). This activity results in the processing activity being subjected to scrutiny by data subjects and supervisory authorities, as required by the GDPR. For controllers already familiar with these tasks, it is a very contained one time effort.
- *Purpose limitation*  
Personal data can only be processed for the purposes for which a legal basis has been found. Art. 29 GDPR describes how this requires instructions for processors and persons who work under the authority of the controller. Such instructions need not be overly detailed and can leave a lot of detail filled in by the expertise and intelligence of these persons. The instructions shall limit the processing to the lawful purposes however. Most organization already have a suitable "line of command" in place, such that the implementation of this requirement requires little effort.
- *Data minimization*  
Only the personal data that is actually necessary to the lawful purposes may be processed by a controller. In the case of *presumed anonymous data* treated as *personal data*, this requires controllers to delete data elements that are unnecessary. This is a very contained one-time effort.
- *Accuracy*  
As reasoned for damage control actions, accuracy is in most cases not expected to be critical for *presumed anonymous data*.
- *Storage limitation*  
By choosing to process *presumed anonymous data*, controllers already comply with this principle and no further action is necessary.

- *Integrity and confidentiality; Security of Processing (Art. 32 GDPR)*  
As in the damage control effort, integrity is not expected to be a major issue here. Confidentiality in contrast is very important here. It likely constitutes the most significant different to treating the data as *anonymous* and also requires the most significant effort.

When keeping *presumed anonymous data* confidential, all disclosure that later cannot be reversed is avoided. This means that the data is only disclosed to intended recipients: Internal recipients who need access to the data in order to satisfy the lawful purposes of the processing; the disclosure to third party recipients requires a valid legal basis and has to make it clear, that the data is considered *personal* and thus requires the protections foreseen in the GDPR.

Confidentiality requires that the data is only accessed for authorized processing (see Art. 5(1)(f) GDPR). Controllers achieve this by implementing technical (such as an access control system) and organizational measures (such as training or non-disclosure agreements). A large part of Art. 32 GDPR “security of processing” is concerned with confidentiality.

Publication of the data, i.e., disclosure to arbitrary third party recipients, is evidently the contrary of confidentiality. Much rather, the disclosure to selected recipient is possible, when there is a valid legal basis for the disclosure. This is the case since according to Art. 4(2) also “disclosure by transmission, dissemination or otherwise making available” is considered to be processing.

Where a controller has direct contact with data subjects before (attempted) anonymization, a commonly used legal basis for the disclosure to third parties is consent. According to Art. 6(1)(a), consent is bound to “one or more specific purposes” of processing. Arguably, these specific purposes cannot be limited to the disclosure itself, but has to include the purposes pursued by the recipients’ processing. This is important to understand that recipients, while bearing their own responsibility for their processing, may have to be restricted in the purposes that they can pursue. Where consent is the legal basis for disclosure, the controller has the responsibility to hold recipients to this limitation.

In any case, the controller disclosing data to third parties must render it clear that the data are considered to be personal data and require the protections afforded to data subjects by the GDPR.

A best practise to propagate the necessary obligations and limitations to recipients is though the stipulation of a legal agreement. This has a similar role as a legal agreement that does the same for processors (see Art. 28(3) GDPR). An example<sup>187</sup> of such an agreement from research practice with pseudonymous (and likely *presumed anonymous*) data is in common use by the Healthcare Cost and Utilization Project (HCUP)<sup>188</sup>. Before the stipulating the contractual agreement, HCUP even vets recipients and requires, among other things, that they pass a test showing that they understand their responsibilities<sup>189</sup>.

---

<sup>187</sup> <https://www.hcup-us.ahrq.gov/team/NationwideDUA.jsp> (last visited 10/5/2021).

<sup>188</sup> <https://www.hcup-us.ahrq.gov/> (last visited 10/5/2021).

- Such a contractual agreement between a controller and a third party recipient could regulate the following:
  - Obligation to treat the data as personal data under the GDPR including implementing measures that guarantee confidentiality;
  - Potentially an obligation to report any breach of confidentiality to the controller;
  - Prohibition of any attempt of re-identification or de-anonymization;
  - Obligation to refrain from further disclosing the data to external recipient or, alternatively, to do so under the same contractual conditions;
  - Potentially the obligation to report any (successful or failed) attempt of re-identification or suitable emerging methodology therefor to the controller;
  - Potentially a limitation of the purposes for which the data can be used (e.g., in the case where the initial disclosure was based on consent);
  - Potentially, where the data permits this, a certain technical protocol for the notifications on the invocations of data subject right invocations according to Art. 19 GDPR.
  - Potentially an obligation to terminate processing and delete the data in presence of any violation of the agreement.
- *Data subject rights (Chapter 3 GDPR)*  
*Presumed anonymous data* are a special case of *irreversibly pseudonymized data without any additional information* (see section 4.9.1 above). According to Art. 11 GDPR, obligations, including data subject rights, that would require a technically impossible identification of the data subject are then waived. A discussion of the waived obligations can be found in section 4.9.7 above). In summary, while there should be a point of contact for inquiries by data subjects, in the case of *presumed anonymous data*, the implementation of right invocation for individual data subject is typically not possible and thus waived.
  - *Data protection by design and by default (Art. 25 GDPR)*  
 When designing the processing operations for personal data, technical and organizational measures in support of the principles of the GDPR have to be taken into account at the earliest possible stage. Considering that identification is presumably already rendered impossible, such measures are typically oriented to preserve confidentiality (see above). To consider them already during the design should not constitute an additional effort.
  - *Processors (Art. 28 GDPR)*  
 Where processors are used for as part of the processing activity of personal data, controllers have to select suitable processors (see Art. 28(1) GDPR) and stipulate a suitable contractual agreement (see Art. 28(3) GDPR). The effort necessary to satisfy this obligation seems contained, particularly if it is based on standard procedures and contracts.

---

<sup>189</sup> See <https://aircloak.com/the-five-private-eyes-part-1-the-surprising-strength-of-de-identified-data/> under HCUP, (last visited 10/5/2021).

- *Data Protection Impact Assessment (DPIA, Art. 35 GDPR)*  
Where a DPIA is not already required, treating *presumed anonymous data* as *personal* does not introduce any additional necessity. In the case where it is required, treating the data as personal renders it easier to demonstrate that the risk to the rights and freedoms of data subjects has been reduced to an acceptable level. In particular, the technical and organizational measures demonstrably support the principles of the GDPR (most prominently, confidentiality).

In summary, the cost of treating *presumed anonymous data* as *personal* is contained. The biggest effort probably lies in the implementation of confidentiality. This also affects how data is being disclosed to third party recipients. Avoiding publication and other forms of disclosure that are not bound to obligations removes the major issue of irreversible actions that was discussed during the damage control effort. Confidentiality and controlled disclosure are thus the most important component of an insurance against incorrect classification of the data.

### **5.7.5 Summary**

In summary, controllers who process *presumed anonymous data* should assess the risk that the data may after all be *personal* now or in the significant time range that needs to be taken into consideration. If this risk is considered to be marginal, it seems like a reasonable approach to treat the data as *anonymous*. Where this risk is above marginal, however, controllers may be well-advised to treat the data as *personal*. This approach results in a contained and plannable effort that insures against potential consequences of a GDPR violation and much more substantial, difficult, and unplannable efforts of the required damage control action.

## **6 Overall summery**

The objective of this document is to clarify some issues around identification, pseudonymization, and anonymization as a basis for writing guidelines about the topic. This document cannot be authoritative, but it attempts to contribute to the present discussion, fostering a better understanding and a consensus of interpretation. It is hoped to also contribute to a future authoritative interpretation.