PARTICIPATORY APPROACHES TO A NEW ETHICAL AND LEGAL FRAMEWORK FOR ICT

**Guidelines on Data Protection Ethical and Legal Issues in ICT Research and Innovation.**

**THE GDPR – MAIN CONCEPTS**

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 788039***.** *This document reflects only the author's view and the Agency is not responsible for any use that may be made of the information it contains.*

# 2 Main Concepts

## 2.1 Personal Data

*Simona Sobotovicova (UPV/EHU)*

*This part of the Guidelines was reviewed by Daniel Jove VIllares, Universidade Da Coruna, Spain*

*This part of The Guidelines has been reviewed and validated by Marko Sijan, Senior Advisor Specialist, (HR DPA)*

### 2.1.1 The concept of personal data

Personal data means any information relating to an identified or identifiable natural person ("data subject"). The definition of personal data under GDPR adds that an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person[1]. There is no doubt that the objective of the rules contained in the GDPR is to protect the fundamental rights and freedoms of natural persons and in particular their right to privacy, with regard to the processing of personal data. However, due to the broad definition of personal data laid down in the GDPR, the Article 29 Data Protection Working Party, the National Data Protection Supervisory Authorities and European Court of Justice (hereinafter, ECJ) case law endorse the definition of personal data.

The Article 29 Data Protection Working Party analysis of the concept of personal data in Opinion 4/2007 has been based on the following four main "building blocks" that can be distinguished in the definition of "personal data"[2]:

- *"Any information"* - This term clearly signals the willingness of the legislator to design a broad concept of personal data. This wording calls for a wide interpretation. It covers "objective" information, such as the presence of a certain substance in one's blood. It also includes "subjective" information, opinions or assessments. Moreover, for information to be "personal data", it is not necessary that it be true or proven.

It must be stated that, the concept of personal data includes a very wide range of information, "not only objective but also subjective", in the form of opinions and assessments, provided that it "relates" to the data subject[3].

- *"Relating to"* - In general terms, information can be considered to "relate" to an individual when it is about that individual. It could be pointed out that, in order to consider that the data "relate" to an individual, a "content" element or a "purpose" element or a "result" element should be present. These three elements (content, purpose, result) must be considered as alternative conditions, and not as cumulative ones, so the presence of one of these elements is enough to be considered to "relate" to an individual.

[1] Article 4(1) GDPR.

[2] *See*, Article 29 Data Protection Working Party: Opinion 4/2007 on the concept of personal data. Adopted on 20th June, 01248/07/EN WP 136, pp.9-12, 21. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf

[3] Judgement of the Court of Justice of the European Union (Second Chamber), Case C-43 4/16, *Peter Nowak v Data Protection Commissioner*, 20 December 2017, §34.

In the words of the EJC the content, purpose or effect criteria act as a parameter for classifying certain information as personal data. If the content, purpose or effect is linked to a particular person, then the information is personal data. The use of one of these criteria is sufficient to exist to classify any given information as personal data[4].

- *"Identified or identifiable"* - In general terms, a natural person can be considered as "identified" when, within a group, this person is "distinguished" from all other members of the group. Accordingly, the natural person is "identifiable" when, although the person has not been identified yet, it is possible to do so (that is the meaning of the suffix "-able").

The GDPR mentions those "identifiers" in the definition of "personal data" in Article 4(1) mentioned previously. Moreover, regarding 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[5]. However, whether the person is "identifiable" is still the focus on the recent scholarly discussions[6].

- *"Natural person"* - The protection applies to natural persons, that is, to human beings. The right to the protection of personal data is, in that sense, a universal one that is not restricted to nationals or residents in a certain country.

The GDPR establishes that natural persons may be associated with online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other identifiers such as radio frequency identification tags. This may leave traces which, in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them[7]. Moreover, the principles of, and rules on the protection of natural persons with regard to the processing of their personal data should, whatever their nationality or residence, respect their fundamental rights and freedoms, in particular their right to the protection of personal data. This Regulation is intended to contribute to the accomplishment of an area of freedom, security and justice and of an economic union, to economic and social progress, to the strengthening and the convergence of the economies within the internal market, and to the well-being of natural persons[8].

[4] Judgement of the Court of Justice of the European Union (Second Chamber), Case C-43 4/16, *Peter Nowak v Data Protection Commissioner*, 20 December 2017, §35.

[5] Recital (26) GDPR.

[6] *See*, for instance; Purtova, N. (2018). The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law. *Law, Innovation and Technology*. DOI:https://doi.org/10.1080/17579961.2018.1452176.

[7] Recital (30) GDPR.

[8] Recital (2) GDPR.

It could be stated that, the Article 29 Data Protection Working Party states that these four elements provided in the first sentence of personal data definition *(any information, relating to, an identified or identifiable and natural person)* are closely intertwined and feed on each other, but together determine whether a piece of information should be considered as "personal data".

### 2.1.2 What information can be considered as personal data?

The National Data Protection Supervisory Authorities and ECJ case law play an essential role in providing interpretation of legal provisions and concrete guidance to controllers and data subjects endorsing a definition of personal data that is wide enough. The definition of the personal data is central element for the application and interpretation of data protection rules which have a profound impact on a number of important issues and topics. Considering the format or the medium on which that information is contained, the concept of personal data includes information available in whatever form, be it alphabetical, numerical, graphical, photographical or acoustic, for example[9]. The ECJ provides a classification of information as personal data in different judgments. To this extent, the term personal data undoubtedly covers the name of the persons in conjunction with their telephone coordinates or information about their working conditions or hobbies. Also information contained in free text in an electronic document may qualify as personal data, provided the other criteria in the definition of personal data are fulfilled. E-mail will for example contain "personal data". The ECJ has spoken in that sense when considering that "referring, on an internet page, to various persons and identifying them by name or by other means, for instance by giving their telephone number or information regarding their working conditions and hobbies, constitutes the processing of personal data [...][10].

On 20 December 2017 the ECJ gave its judgment on the "*Nowak* case"[11] establishes the classification of the answers and subjective comments of the examiner within the written answers submitted by a candidate in a professional examination as personal data, establishing a series of criteria that make it possible to understand which data are of a personal nature[12]. The ruling addresses the potential application of GDPR to constitute personal data[13]. It must be highlighted that, the classification of this data as

[9] Article 29 Data Protection Working Party: Opinion 4/2007 on the concept of personal data. Adopted on 20th June, 01248/07/EN WP 136, p.7. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf

[10] Judgment of the European Court of Justice, C-101/2001, *Lindqvist*, §27, 06.11.2003.

[11] Judgement of the Court of Justice of the European Union (Second Chamber), Case C-43 4/16, *Peter Nowak v Data Protection Commissioner*, 20 December 2017.

[12] Jove, D. (2019). Peter Nowak v Data Protection Commissioner: Potential Aftermaths Regarding Subjective Annotations in Clinical Records. *European Data Protection Law Review*, Volume 5, Issue 2, p. 175. DOI: https://doi.org/10.21552/edpl/2019/2/7

[13] Judgement of the Court of Justice of the European Union (Second Chamber), Case C-43 4/16, *Peter Nowak v Data Protection Commissioner*, 20 December 2017, §27.

personal data entails, for the candidate, the possibility of using their rights of access, rectification and objection. To this extent, the classification as personal data provides the right of access, but also the other powers given to the owner of this type of data, which are: rights of rectification, erasure and objection, as well as all the guarantees included in the data protection legislation[14].

The sentence also analyzes the applicability of the right of access to data with more than one owner and opposing interests (in this case the examiner and candidate). The ECJ reaffirmed the idea that, the fact that the information is in the hands of one person or several people is irrelevant regarding its classification as personal data. The attribution of the condition of personal data does not come from this fact, but from the very nature of the information. Regarding the definition of personal data, the ECJ adds another feature to this: the plurality of affected persons, or the possibility that one piece of information may be personal data of more than one data subject[15].

Due to the classification of an information as personal data, in the *YS and Others*[16] case, it is considered that the legal analysis of a minute produced within the framework of a request for a residence permit, is not personal data as it refers to "information about the assessment and application by the competent authority of the law to the applicant´s situation. This interpretation meant that, in the *YS and Others* case, the right of access was not recognized for that information, believing that such access would be based on a right of access to public documents which is not covered under GDPR legislation[17]. However, if the analysis had included any evaluations of the subject, or that could have an effort on them, then this would be considered as personal data which would, as such, be subject to the GDPR[18].

It could be affirmed that, the GDPR definition, as recalled by the ECJ, is based on the broad definition of personal data reflecting the intention of the legislator to assign a wide scope to the concept, encompassing subjective and objective information on data subject. Since the classification of information as personal data brings it into the realm of the fundamental rights protection architecture of the EU, it also establishes both the rights of the data subjects and the circumstances under which the standard of protection may be diminished due to justifiable objectives[19].

[14] Jove, D. (2019). Peter Nowak v Data Protection Commissioner: Potential Aftermaths Regarding Subjective Annotations in Clinical Records. *European Data Protection Law Review*, Volume 5, Issue 2, p. 177. DOI: https://doi.org/10.21552/edpl/2019/2/7

[15] *Ibídem*, p. 176, 178.

[16] Judgment of the Court, Joined Cases C□141/12 and C□372/12, *YS and Others*, 17 July 2014.

[17] Judgment of the Court, Joined Cases C□141/12 and C□372/12, *YS and Others*, 17 July 2014, §40.

[18] Jove, D. (2019). Peter Nowak v Data Protection Commissioner: Potential Aftermaths Regarding Subjective Annotations in Clinical Records. *European Data Protection Law Review*, Volume 5, Issue 2, p. 179. DOI: https://doi.org/10.21552/edpl/2019/2/7

[19] Podstawa, K. (2018). Peter Nowak Data Protection Commissioner: You can access your exam script, because it is personal data. *European Data Protection Law Review (EDPL),* 4(2), pp. 254, 256. DOI: https://doi.org/10.21552/edpl/2018/2/17.

## 2.2 Data Processing

*Iñigo de Miguel Beriain (UPV/EHU)*

*This part of the Guidelines was reviewed by Daniel Jove Villares, Universidade Da Coruna, Spain.*

*This part of The Guidelines has been reviewed and validated by Marko Sijan, Senior Advisor Specialist, (HR DPA).*

### 2.2.1 Definition

According to article 4(2) of the GDPR, processing "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction".

Therefore, the concept of processing is broad. It covers a wide range of operations performed on personal data, including by manual or automated means, if it is part of a structured filing system, that is, a structured set of personal data which are accessible according to specific criteria, whether centralized, decentralized or dispersed on a functional or geographical basis (art. 4(6)).

Clearly, the list included in article 4(2) is non-exhaustive, meaning that other operations with personal data that work well with the general definition should also be considered processing under the GDPR. Some examples of processing include: staff management and payroll administration; access to/consultation of a contacts database containing personal data; sending promotional emails; shredding documents containing personal data; posting/putting a photo of a person on a website; storing IP addresses or MAC addresses; video recording (CCTV), etc. [20]

### 2.2.2 Processing as a key concept in the GDPR

Processing is an essential element in terms of data protection rights. What the GDPR really regulates is not the data itself, but the processing of personal data. This use of data triggers the application of data protection regulations. Indeed, article 1(1) of the

[20] EU Commission, What constitutes data processing, at: https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-constitutes-data-processing_en.

GDPR states that “This Regulation lays down rules relating to the protection of natural persons **with regard to the processing of personal data** and rules relating to the free movement of personal data.”

The circumstances of the processing define the essential regulatory elements: the need (or not) to find a reason to process the data, if it is of a special category; the appropriate basis of legitimacy; whether it is a one-to-one treatment or processing on a large scale; the specific level of risk; the safeguards to be implemented; and so on. Each processing will be, in short, a separate, independent event, with its own characteristics and scale. Hence, it is always necessary to think that data protection regulations apply to each of them.

## 2.3 Data Protection by Design and by Default

*Bud P. Bruegger (ULD)*

*Acknowledgements: The author thankfully acknowledges Kirsten Bock’s help with legal interpretation, Harald Zwingelberg’s feedback and review and a detailed review and suggestions by Hans Graux*

*This part of the Guidelines was finally validated by Hans Graux, guest lecturer on ICT and privacy protection law at the Tilburg Institute for Law, Technology, and Society (TILT) and at the AP Hogeschool Antwerpen. President of the Vlaamse Toezichtscommissie (Flemish Supervisory Committee), which supervises data protection compliance within Flemish public sector bodies.*

The present section attempts to provide practitioners with a more detailed understanding of how to practically implement the requirements of Art. 25 GDPR *Data Protection by Design and by Default (**DPbDD**).*

The present section on DPbDD is structured as follows:

A first subsection discusses the guidelines on the topic issued by the EDPB. It points out the differences to the approach taken here.

A second subsection describes the scope of the obligations arising from Art. 25 GDPR. Most importantly, it clarifies in which way technology providers are affected by it.

A third subsection analysis Art. 25 GDPR. Since Art. 25(1) mandates controllers to implement measures both at the *time of determining the means* and at the *time of processing itself*, the precise meaning of ***determining the means*** and ***processing itself*** is discussed. This relies on an analysis of what the GDPR states about the structure of processing. The analysis of Art. 25(1) also puts emphasis on the meaning of

***effectiveness*** of measures. The discussion of Art. 25(2) explains what exactly is meant by the term ***default*** and analysis the obligations of the controller.

A fourth subsection focusses on the actual processes that implement data protection by design. In particular, it describes the processes to implement DPbDD in the three main phases of *determining the purposes*, *determining the means*, and the *processing itself*. These processes aim at a systematic implementation of the data protection principles in every work task of each phase. This then results in the identification and implementation of technical and organizational measures.

### 2.3.1 Guidelines by the European Data Protection Board

The European Data Protection Board (EDPB) has issued guidelines on Data Protection by Design and by Default[21]. It emphasizes the importance of understanding and applying the **data protection principles** (see the "Main Principles" section in the general part of these Guidelines) and of implementing **data subject rights** (see the "Data Subject Rights" section in the general part of these Guidelines).

The importance of the data protection principles is for example expressed in paragraph 61: "Controllers need to implement the principles to achieve DPbDD. These principles include: transparency, lawfulness, fairness, purpose limitation, data minimization, accuracy, storage limitation, integrity and confidentiality, and accountability. These principles are outlined in Article 5 and Recital 39 of the GDPR. To have a complete understanding of how to implement DPbDD, the importance of understanding the meaning of each of the principles is emphasized."

The importance of data subject rights is stated in paragraph 63: "While this section focuses on the implementation of the principles, the controller should also implement appropriate and effective ways to protect data subjects' rights, also according to Chapter III in the GDPR where this is not already mandated by the principles themselves."

The EDPB guideline dedicates its section 3 to the implementation of data protection principles. The PANELFIT guidelines go beyond this by providing a more detailed description of each principle together with many examples of technical and organizational measures suitable to implement those principles.

Like the EDPB guidelines, the following text also analyzes the meaning of Article 25 GDPR. The present text attempts to provide additional concrete guidance, however. To achieve this, it not only provides a legal analysis of the phases of processing according to the GDPR, but also provides a technical analysis of what tasks are necessary for each phase. In particular, this is done for *determining the means of processing* and for *the processing itself*. In each of the tasks that are identified, the data protection principles can then be applied and technical and organizational measures identifies and implemented.

[21] European Data Protection Board, Guidelines 4/2019 on Article 25 Data Protection by Design and by Default, Version 2.0, Adopted on 20 October 2020, https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf (last visited 30/11/2021).

A second major difference from the present text and the EDPB guidelines is that the former discusses the actual process necessary for applying DPbDD in the various phases.

A minor difference is that the present text goes into further detail on how controllers can pass on requirements to producers of software and services. The text does not go into the merit of certification, however; should this be relevant to readers, they are referred to the EDBP guidelines.

### 2.3.2 The scope of DPbDD

This section discusses how the GDPR contains solely obligations for controllers (and processors) and how this can indirectly influence technology providers.

Data protection by design can be seen as taking data protection into consideration not only for *processing operations* taking place in the operational phase, but also earlier in the planning and implementation phases. More generally, one could see data protection by design as a methodology that takes data protection into account in all phases of the life cycle of a *processing activity*[22], ranging from its conception, over design and implementation, to operational use and final dismantling.

The complete life cycle typically involves activities by players other than the controller and processor. Most importantly, many decisions that affect data protection aspects of a processing activity are taken by technology providers, who often design and implement software and systems. Where technology providers invest in developing products and services that are then offered on the market, they also contribute to defining the *state of the art* of a certain type of processing of personal data.

In contrast, the GDPR expresses obligations for controllers and processors. It lacks any direct obligation for technology providers. In its *Preliminary Opinion on privacy by design*[23], the EDPS points out this fact by stating the following[24]:

“A serious limitation of the obligations of Article 25 is that they apply only to impose an obligation on controllers and not to the developers of those products and technology used to process personal data. The obligation for products and technology providers is not included in the substantial provisions of the GDPR.”

Since the GDPR as a whole, and Art. 25 in particular, express solely obligations for controllers (and processors), the scope of the present section is limited accordingly.

While there are no legal obligations for technology providers, Art. 25 GDPR nevertheless influences them indirectly. Recital 78 GDPR hints at this by stating the following[25]: *“The principles of data protection by design and by default should also be*

[22] The term *processing activity* is here used in the sense of Art. 30 GDPR *records of processing activities* and 4(16)(b) GDPR. In both cases, a *processing activity* is the basic unit of undertaking by a controller that involves the processing of personal data.

[23] The European Data Protection Supervisor (EDPS), Opinion 5/2018, Preliminary Opinion on privacy by design, 31 May 2018, https://edps.europa.eu/sites/edp/files/publication/18-05-31_preliminary_opinion_on_privacy_by_design_en_0.pdf (last visited 29/6/2020).

[24] Pages 7 and 8, Side number 37.

[25] See sentence 5.

*taken into consideration in the context of public tenders."* How the influence on technology providers happens is described in more detail in the sequel.

The argument focusses on software that is created by a technology provider. There are two options for how a controller can obtain such software:

- As the result of a custom development, or
- By acquiring the software on the market.

In the former case, the software house, the design and development is triggered by the controller and the technology provider can be either internal or external; in the latter case, there is a multitude of controllers with similar needs who create a market demand for certain kinds of software. The design and development of the software is then triggered by the technology provider with the objective of achieving a competitive position in the market.

The technical details inherent in software development are usually inaccessible to controllers and their representatives. Therefore, in both cases, the interaction between controllers and technology providers is limited to communication about requirements. In particular, the role of requirements in the two cases is as follows:

- In the case of custom development, requirements are the main tool for controllers to express the objectives of the development process. The requirements are also used to determine whether the development process has successfully terminated. This happens during acceptance testing.
- In the case of controllers buying software, they need requirements to guide their selection of adequate software from the offering of the market. In tenders, such requirements can be communicated to technology providers in order to solicit offers that are adequate for the needs; where software is bought without tender, controllers must verify whether various candidate software offerings satisfy the requirements. In both cases, the validation of offerings relative to the requirements are a major factor in the purchasing decision by the controller.

So while obligations for technology providers are out of scope of Art. 25, controllers are obliged to determine adequate data protection requirements and bear the full responsibility for the software they operate. The validation of software against requirements can take into account the state of the art and the cost of implementation (see Art. 25 GDPR and discussion later). The absence or excessive cost of adequate software on the market cannot be considered a valid justification for operating inadequate software, however.

### 2.3.3 Analysis of Article 25. Data protection by design

The present section analyzes the letter of the law with the **objective** of finding **a structured and systematic approach** to discuss the measures that controllers are mandated to implement by Art. 25 GDPR. The resulting systematics and structure are then **used in section 2.3.3.1 on measures** which constitutes the most concrete guidance for practitioners.

To foster clear understanding of the text, the following breakout box defines two often used terms.

Definition: ***processing activity***

The term *processing activity* is here used in the sense of Art. 30 GDPR *records of processing activities* and 4(16)(b) GDPR. In both cases, a processing activity is the basic stand-alone unit of undertaking by a controller that involves the processing of personal data. A processing activity undergoes a life cycle that includes conception, design, implementation, operation, and dismantling.

Definition ***processing operation***

The term *processing operation* refers to only the operational phase of a *processing activity* where a processing system is operated to actually process personal data. It entails the execution of *processing operations* as they are defined in Art. 4(2) GDPR. Other aspects of processing activities, such as conception and design, fail to execute such processing operations and are therefore not deemed part of the processing operations.

#### 2.3.3.1 Overview and main obligation for controllers

Art. 25 GDPR includes the following:

Art. 25(1):

*Taking into account [..],* ***the controller shall****, both at the time of the determination of the means for processing and at the time of the processing itself,* ***implement appropriate technical and organizational measures*** *[..] which are designed to implement data-protection principles [..] in an effective manner and to integrate the necessary safeguards into the processing [..].*

The main obligation for controllers stated in Art. 25(1) GDPR is thus that they "***shall*** *[..]* ***implement appropriate technical and organizational measures*** *[..] which are designed* ***to implement data-protection principles***" (see the "Main Principles" section in the general part of these Guidelines).

Throughout the GDPR[26], the implementation of technical and organizational measures is stated to be the way to comply with data protection principles. This implies that everything a controller does in support of the data protection principles must be considered to be a measure. Consequently, the **concept of measure** must be **understood in a very broad sense**. This means that it is not restricted to physical artefacts (such as firewalls), or specific actions (such as training of staff). It must also encompass all considerations and decisions that are necessary to determine the means of processing in a manner that is compliant with the principles and obligations of data protection.

Art. 25(1) GDPR also states that these measures shall be implemented "***in an effective manner***". Efficiency will therefore be analyzed below.

[26] This includes among others Art. 24, 25, and 32 GDPR.

Furthermore, Art. 25(1) states that the measures are implemented "***to integrate the necessary safeguards into the processing***". In other words, the implementation of measures is the way to achieve the objective of integrating the necessary safeguards into the processing. Grammatically, this interpretation becomes even clearer when expanding "*to integrate*" into its complete form of "*in order to integrate*". The "to" excludes the interpretation that, in addition to the *implementation of measures*, also *the integration of safeguards* is required.

Arguably, the essence of Art. 25(1) lies in the wording "*both* ***at the time of the determination of the means*** *for processing and* ***at the time of the processing itself***". This means that the implementation of the measures has to happen in **two distinct periods of time**. It thus implies a **phase-model for a processing activity**. This is compatible with the understanding of data protection by design, as considering data protection in every phase of a processing activity. The legal interpretation of the phases of processing addressed in Art. 25(1) is provided in the following subsection.

#### 2.3.3.2 The phases of processing in the GDPR

Art. 25(1) GDPR speaks of two phases in relation to a processing activity, namely "**the *time of the determination of the means* for processing**" and "**the *time of the processing itself***". It is evident that both these *times* must be time periods of a certain duration rather than points in time. It is also evident that the time of determining the means must precede the time of the processing itself. We therefore call these time periods also ***phases***.

Art. 4(7) states that in addition to the means, the controller also "**determines the purposes**". This evidently also takes time and precedes the determination of the means. It seems useful to include the determination of purposes for completeness and in case there are measures that can be implemented in that phase.

Consequently, the GDPR implies the following **phase model of a processing activity**:

- Phase 1: Determination of the purposes;
- Phase 2: Determination of the means;
- Phase 3: Processing itself.

To better understand what exactly happens in each phase, it is necessary to analyze in more detail what the GDPR defines as a processing operation.

#### 2.3.3.3 Processing operations in the GDPR

The following analyzes what the GDPR defines as a processing operation.

Art. 5(1)(f) GDPR states the necessity of "protection against unauthorized [..] processing". This implies that the ordinary processing needs to be **authorized**. It is also clear from the context that such authorization must come from the controller who bears the full responsibility for the processing. But how can a controller limit the processing to what is authorized?

A partial answer to this question can be found in Art. 29 GDPR: "The processor and **any person acting under the authority of the controller** or of the processor, who has access to personal data, shall not process those data except **on *instructions* from the controller**, [..].” Also Art. 32(4) GDPR uses a very similar wording. Art. 29 GDPR implies the following conception:

- The processing operation is executed by a “**natural person acting under the authority of the controller or the processor**”. Such persons are most often *employees* of the controller, but could also work for a processor or work without actual employment[27]. They are called ***human resources*** in the sequel. Note that these persons in turn control technical means that support or partially automate the processing[28].
- The **means** with which a **controller ensures that only authorized processing takes place** is through issuing ***instructions***.

To ensure that only authorized processing takes place, the instructions must specify all relevant aspects of the processing activity: **who, when, what,** and **how**. In other words, human resources need to act only **on instruction** (who, when) and **as instructed** (what, how).

Albeit with less clarity, the GDPR also states that **technical resources** are necessary. This is very clear in Recital 39 (sentence 12) that speaks of the “**equipment used for the processing**”. Other terms related to technical resources that are used in the GDPR are “data processing equipment” in Art. 58(1)(f) and “processing systems” in Art. 32(1)(b).

While the GDPR uses the term *instruction* only in the context of human resources, it is clear that **also technical resources require instructions** in order to execute only authorized processing. In the technical domain, the term *machine instructions* is used here. An important type of such instructions is ***software***.

In summary, when looking at an **individual** (human or technical) **resource**, the GDPR defines a processing operation as follows:

**individual processing operation**
=
**execution of** the controller’s **instructions by a** single **resource**

In most cases, the **overall processing operations** involve a system of a multitude of interacting human and technical resources. This is expressed in the following:

**overall processing operations**
=

[27] See also EDPB guidelines on the concepts of controller and processor in the GDPR, paragraph 88 for a discussion of the meaning of “persons who, under the direct authority of the controller or processor, are authorized to process personal data”.

[28] Note that even in the case of “fully automatic processing”, it is always a person who controls such processing by starting and stopping it. The control by a person is even more evident when looking at computerized “tools” that are used by humans though a human-machine-interface.

**multitude of** individual **processing operations**
executed **by individual** human and technical **resources**

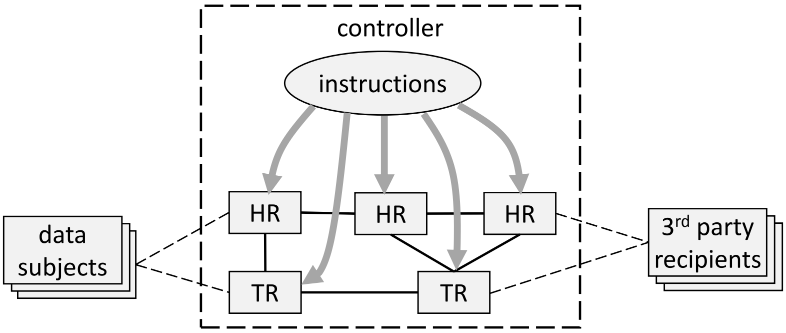


Figure 6: The GDPR's conception of a processing operation.

Figure 6 illustrates the GDPR’s concept of processing operations in a wider context. It illustrates the domain of responsibility of the controller by a dashed box. The controller determines the authorized processing operations by issuing or selecting/approving[29] instructions to both, human resources (HR) who act under its authority (see Art. 29 GDPR) and technical resources (TR) under its control. All resources interact to form the overall processing system. The context of this processing system is defined by *data subjects who* interact with human and/or technical resources, and optionally *third-party recipients* (see Art. 4(9) and (10) GDPR) to whom resources disclose personal data.

This model of processing operations represents the processing authorized by the controller. It is used in the next section to better understand what *determining the means* actually entails.

#### 2.3.3.4 Determining the means

In its guidelines on the concepts of controller and processor in the GDPR[30], the European Data Protection Board provides a legal analysis of what it means to *determine the means* of processing. The discussion here is more technically oriented. The Board distinguishes between “essential” and “non-essential means”; the latter can also be determined by processors. The present text does not make such a distinction and just provides a technical interpretation of what decisions the determination of the means entails.

*Determining the means* is a phase that predates the operational use of a processing system and prepares and sets up everything that is necessary for the actual processing operations. This means, that guided by the *purposes*, a controller has to plan, design, and implement everything necessary to enable the *processing itself*. This includes at least the following tasks:

- **Determine** the human and technical **resources** necessary for the processing;

[29] Selecting and approving of instructions by a controller is for example done when off-the-shelf software is acquired or when a controller chooses the service of a given processor.

[30] European Data Protection Board, Guidelines 07/2020 on the concepts of controller and processor in the GDPR, Version 2.0, Adopted on 07 July 2021, https://edpb.europa.eu/system/files/2021-07/eppb_guidelines_202007_controllerprocessor_final_en.pdf (last visited 2/12/2021).

- **Determine** the **instructions** that **define the authorized processing** and are suitable for the resources;

What this entails in more detail is described in the following.

**Determining human resources** entails at least the following:

- Planning that determines what human resources are necessary, selecting suitable human resources and bringing them under the authority of the controller or processor. This is typically done through employment that establishes a contractual relationship between the employee and the controller.
- Setting the human resources in a condition in which they can translate the instructions in a manner that constitutes authorized processing. This can entail things such as
  - the undersigning an agreement to refrain from disclosing personal data,
  - the commitment by the human resource to certain general policies or codes of conduct, and
  - training of the human resource to acquire knowledge and skills necessary for executing the instructions in the desired way.

**Determining technical resources** entails at least the following:

- Planning, selecting and acquiring the necessary technical resources.
- Bringing the technical resources in a condition that they can execute the necessary processing operations. This can entail things such as
  - physical installation,
  - configuration,
  - integration in the used infrastructure, and
  - installing the necessary software.

**Determining instructions for resources in general:** After having discussed the determination of resources, the following analyzes the determination of instructions. Looking at Figure 6, it is clear that the following kinds of instructions exist:

- Instructions that determine the behavior of a **single resource**,
- instructions that determine the **interaction between** multiple **resources**,
- instructions that determine the **interaction** between resources and **data subjects**, and
- instructions that determine the disclosure of personal data to **third-party recipients**.

These different types of instructions are discussed in more detail in the following while distinguishing between human and technical resources.

**Determining the instructions for human resources** are discussed in the following:

- **Instructions for individual human resources** can be expressed in two different styles:
    - Defining the required outputs, products, and effect the activity of the human resource shall result in. This constitutes ***declarative instructions*** that focus on the ***what*** aspect and relying on the resource's capability to fill in the *how* aspect of the instructions.
    - Detailed descriptions of the way in which an activity has to be executed. This constitutes ***imperative instructions*** that focus on the ***how*** aspect, require less intelligence and autonomy from the executing resource, and often define the *what* aspect in a more implicit way.
- **Instructions on how human resources interact among each other**: This includes designing and specifying *business processes*, *work-flows* and *data-flows*. There are various formal languages[31] and graphical notations[32] to support such activities.
- **Instructions on how human resources interact with technical resources**: Technical resources are not autonomous. They are controlled by humans (i.e., human resources). Even the most autonomous technical resource needs to be switched on. Usually, human resource exercise a further-reaching control over the technical resource through ***user interfaces*** and via *human machine interaction* (HMI). A common way to model such interactions are ***use case diagrams***. These are often also used for the *specification* of *functional requirements* of software. Instructions on how humans interact with technical resources also define which human resources are **authorized to access** which technical resources for what purposes. These kinds of instruction thus also determine the **responsibility** human resources have for operating certain technical resources.
- **Instructions on how human resources interact with data subjects** determine what interactions data subjects can have with the controller. This includes the manual processing of data subject right invocations (see Chapter 3 GDPR) and foreseen interactions with the data protection officer (see Art. 38(4) GDPR).
- **Instructions on how human resources interact with third-party recipients** determine which personal data is manually disclosed to third-party recipients.

**Determining the instructions for technical resources** entails the following:

- **Instructions for individual technical resources** potentially encompasses the following aspects:

[31] These formal languages include for example the *XML Process Definition Language* (XPDL) and the *Business Process Execution Languag*e (BPEL).

[32] These graphical visualizations include for example the *Business Process Model and Notation* (BPMN), *Activity Diagrams*, *flow charts*, and *Petri Nets*.

  - Procurement[33] of ***Software*** which usually constitutes machine instructions which are expressed in some formal (imperative or declarative) programming language. The behavior of the software may depend on parameters that can be determined at a later point of time; such parameters are typically called *configuration*. The decision whether configuration is possible and what parameters it entails is built into the software. There are two types of configuration,
    - the one determined by the controller, and
    - the one controlled by the data subject (for example *preferences* and *settings* supported by an appropriate user interface).
  - ***Configuration*** of the software **by the controller**.
  - **Specification of *default values*** for configurations performed by the data subject. This is obviously the subject of ***Data Protection by Default*** that is regulated in Art. 25(2) GDPR (see section 2.3.4 below).
- **Instructions on how technical resources interact among each other:** Technical resources can interact with each other when they have ***interfaces*** that are connected by communications ***channels***. The communications that can take place are typically determined by ***protocols***. Communications can be represented, for example, by ***interaction diagrams*** such as *UML sequence diagrams* and *UML communication diagrams*. Such communications typically involve the exchange of (personal) data. These can be represented graphically in ***data flow diagrams***. This kind of instructions also determines which technical resources are **authorized** to interact with which others and for what purposes. The aspects determined by these kinds of instructions are often related with the concept of ***technical (component) architecture***.
- **Instructions on how technical resources interact with data subjects** are typically used for **configuration by data subjects** (see Art. 25(2) GDPR and its discussion below) and for the automated support of ***data subject rights*** (see Chapter 3 GDPR). Both require ***appropriate user interfaces***. Again, *use case diagrams* can be used to represent these. Again, ***authentication*** and ***access control*** is necessary for the technical resource to determine whether the user is indeed the claimed legitimate data subject.
- **Instructions about automatic transfer of data to third-party recipients** determine which (personal) data is disclosed, under which conditions, and how. This typically requires interfaces for humans or machines and appropriate channels of communications. Data can be pushed to recipients or disclosed on request. Authentication (of humans or machines) and access control are typically relevant also here.

[33] Procurement is here used as a collective term that encompasses both, custom development and acquisition of software from the market. In both cases, controllers are responsible for an adequate requirements analysis and specification.

**Identifying appropriate technical and organizational measures:** As was discussed in section 2.3.3.1 above, Art. 25(1) mandates controllers to implement *appropriate technical and organizational measures which are designed to implement data-protection principles* also at the time of *determining the means*. It has been reasoned above, that determining the means consists of determining the resources and the instructions. Also, together, resources and instructions constitute a processing system able to execute the authorized instructions on actual personal data.

It is clear that the required measures must be integrated with this processing system. Namely, they must be integrated into its instructions and applied to its resources. In other words, such measures cannot be determined independently. In fact, they need to be determined together with the determination of instructions and resources. In every step of determining a part or aspect of the processing system, the principles of data protection have to be taken into account in order to identify and integrate adequate measures.

For this reason, the aspects of a processing system that were distinguished in the above discussion directly identify the areas where appropriate measures have to be found and implemented. This section is therefore instrumental in providing a structure for the detailed discussion of measures in section **2.3.3.1 above**. It also serves to achieve a certain completeness by systematically considering all aspects and each principle.

### 2.3.3.5 Processing itself

Processing itself is **started** by the **go-ahead** from the controller to the resources to start executing the issued instructions. From this point on, the **processing of actual personal data** starts to take place. Namely, it is executed by the designated resources who follow the controller's instructions.

The *processing itself* **terminates** when no more personal data are being processed. Considering that according to Art. 4(2) GDPR the *storage* of personal data constitutes processing, termination of *processing itself* goes beyond just telling resources to stop executing the issued instructions. It also requires **additional instructions** to ascertain that personal data is not being stored any longer. We call this ***dismantling*** of the processing operations. Dismantling encompasses ***erasure*** and ***destruction*** of personal data, both of which still constitute *processing* according to Art. 4(2).

Art. 25(1) requires controllers to implement also appropriate technical and organizational measures during the processing itself. In analogy to the *determination of means*, the structure found for the processing itself will be used to guide the discussion of measures.

### 2.3.3.6 Re-determining the means during operational processing

Considering that the result of determining the means are the resources and the instructions, it is common place to re-determine the means also during operational processing. The following examples shall illustrate this:

- **Replacement of** failing technical **resources** and unavailable human resources. Replacement of resources can be temporary or permanent.

- **Addition, subtraction or replacement of resources** to adapt to a **changing volume of processing**. This could include, for example, the addition of human resources to an overtaxed work unit or the replacement by a technical resource by a more powerful one.
- **Change of instructions** for improved **efficiency and effectiveness**. This can include, for example, routine updates of software to the latest version. Other examples are the evolutionary improvement of instructions or the redesign of organizational processes.
- Beyond this, also an **extension of the means** to support an **extension of the purposes** is possible. This typically goes along with additional functionality supported by the processing.

Since such re-determining the means is still determining the means, also here, appropriate measures have to be implemented by the controller. The above analysis will therefore also be used for structuring the discussion of means in section 2.3.3.1 above.

#### 2.3.3.7 Effectiveness of measures

The following analyzes the requirement of Art. 25(1) GDPR that the measures need to be implemented "***in an effective manner***". It does so in the context of the other wording of Art. 25(1) GDPR.

Unlike the previous analysis, the present one will not be used to identify areas for which measures have to be found. It will be used as an important aspect that needs to be considered for each of the proposed measures.

Art. 25(1) GDPR mandates controllers to implement appropriate measures "*which are designed to implement* ***data-protection principles***" in order "*to integrate the necessary* ***safeguards*** *into the processing*". In this context, the requirement of effectiveness expresses that it is not an objective in its own right to implement measures. In fact, measures are only of value based on their **effectiveness to implement the data-protection principles** and **to integrate safeguards**. Consequently, just implementing measures without considering their effectiveness would be a futile exercise.

The contexts relative to which effectiveness has to be analyzed are provided in Art. 25(1) GDPR in form of the aspects which controllers need to take into account. Namely, these aspects are the following [listed in a different order than used in the text of the GDPR]:

- "*the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing*",
- "*the cost of implementation*",
- "*the state of the art*", and
- "*the nature, scope, context and purposes of processing*".

When considering effectiveness in the **contexts of the risks** to affected natural persons, it is evident that the measure must be effective to mitigate the risks. It also implies a certain proportionality relative to the magnitude of the risks. When considering a **set of**

**implemented measures**, their effectiveness is **sufficient** if it is suited to **mitigate the risk to an acceptable level**.

When considering the effectiveness in the **context of cost**, the GDPR seems to acknowledge that the resources available to implement measures are limited and should be used effectively. This permits controllers to use less expensive, cost-effective, measures in place of expensive ones with a similar effect. In other words, the criterion is effectiveness, not affordability or cost to the controllers as such. While the consideration of cost leaves the possibility that a cost can be deemed excessive, high cost cannot be used as a justification to disregard the effectiveness required in different contexts. If the costs required to ensure an adequate level of guarantees are too high for a controller, the controller should refrain from the processing activities.

When considering the effectiveness in the **context of the state of the art**, the consequences are two-fold. On one hand, it prevents controllers from ignoring new measures and refraining to update the level of protection to what is offered by the state of the art. On the other hand, a controller cannot be obliged to implement measures that have been outlined in some research paper without having been tested or rendered usable in an operational environment. In situations where controllers rely on the market to provide certain kinds of software, controllers may be justified to limit the implemented measures to those actually available on the market, if these are sufficient to provide effective protections. As in the context of cost, this cannot waive effectiveness requirements in other contexts, however.

In the context of security measures, the state of the art has a particular meaning. Cybersecurity can be seen as an arms race between attackers and defenders. In the ever evolving threat landscape, whenever defenders mind more effective means of thwarting attacks, attackers find more sophisticated means of attack. This makes it evident that the concept of an "effective defense" is constantly moving. In this context, current information about threats and available defenses are important when assessing the effectiveness of implemented measures. Also, a failure to implement new measures, for example in the form of security-critical updates or patches, cannot be justified by controllers (except in the rare event where the new measures are irrelevant to the processing activities and the related risks).

Note that the EDPB points out in their guidelines on DPbDD that the state of the art is not only defined by technical measures, but also includes organizational measures such as frameworks, standards, certification, and codes of conduct[34].

When considering the effectiveness relative to the **nature, scope, context and purposes of processing**, it is acknowledged that measures have to be matched with the processing at hand. A measure that is effective for a traditional information system that supports humans who make decisions may not be effective when applied to a machine-learning application that makes automatic decisions; a measure that works fine for low-volume processing in a small environment may not scale up to high-volume processing; and a measure that works effectively when using trustworthy processors (whom themselves are subject to the GDPR) may not be effective and sufficient when using

[34] See paragraph 22 of EDPB guidelines on DPbDD.

less trustworthy processors (such as those located in 3rd countries and not themselves bound by the GDPR).

Art. 5(2) requires that controllers must be able to demonstrate compliance with the GDPR. An important aspect of this is to be able to demonstrate that the implemented measures are indeed effective. It should be an integral part of the process of making decisions about which measures to implement. The dimensions of effectiveness are given in Art. 25(1) and have been discussed above.

### 2.3.4 Analysis of Data Protection by Default in Art. 25(2) GDPR

The following will analyze the requirements of Art. 25(2) GDPR. It uses the definition of defaults provided in the *determining the instructions for technical resources* section in this document (1.3.3).

As clear from the above definition, defaults pertain to settings (sometimes arranged as *preferences* or *user profile*) that are under the control of the data subject. Controllers decide about the **default settings**, i.e. the **settings** that are active **in the absence of any intervention on part of the data subject**.

These settings influence the processing that takes place, including the following aspects:

- the personal data that are being processed,
- the extent of processing that is performed,
- the period for which the data are stored, and
- the natural persons to which the personal data is made accessible.

The following example of settings shall illustrate this:

- Data subjects can optionally provide an **e-mail address** in order to **be informed about the processing status of an order**. Evidently, this affects the amount of personal data that is processed by the controller. It also affects the extent of processing.
- For an order processing, data subjects always have to provide a s**hipping address and payment information**. Optionally, they can click a box to **remember** this information to **avoid typing it in repeatedly** for future orders. While the amount of data processed by the controller is always the same, the user-controlled option obviously affects the storage period of that data.
- A **social media** provider may present its users with **privacy settings** that control the **visibility of their posts**, ranging from *only close friends* to *everybody.* Evidently, this privacy setting controls the natural persons who have access to the posts, which represent personal data.

The GDPR includes the following:

> Art. 25(2):
>
> *The controller shall implement appropriate technical and organizational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of*

> *personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.*

Art. 25(2) thus mandates that **by default**, the processing shall be **limited** to what is **necessary for the purposes**. It further clarifies that this must be understood in respect of the **amount of data**, the **extent of processing**, and the **period of data storage**. The third sentence states that this is also applicable[35] to the number of persons to which the data is made accessible. This thus seems to refer to the number of recipients (as defined in Art. 4(9) GDPR).

The wording of Art. 25(2) implies that there must be some kinds of additional purposes: by default, the processing must be limited to a certain set of purposes; but after the intervention of the data subject, evidently the processing goes beyond this limitation. This implies, the processing then pursues additional purposes.

The above examples help to understand this better. In the first example, the additional purpose is to **keep the data subject informed about the processing status** of orders. In the second example, the additional purpose is to improve **user convenience** for those data subjects who expect to place orders again in the future. In the third example, no additional purpose is pursued. In fact, the purpose of restricting the visibility of social media posts to the **range intended by the user**, is always present. Note that the third sentence of Art. 25(2) that fits this example also refrains from making reference to purposes.

These examples illustrate that the **additional purposes** and the purposes underlying the situation addressed in the third sentence are always **purposes that benefit the data subjects**.

Based on this analysis, Art. 25(2) seems to state that **by default**:

- **additional purposes** that may benefit data subjects shall be **disabled**, at least as long as they require the collection additional data, increase the extent of processing, cause an extension of the storage period, or increase the number of recipients;
- where a purpose in the interest of the data subject is always pursued by the processing (i.e., cannot be disabled), its **data protection impact must be minimized** regarding collected data, extent of processing, storage period, and number of recipients.

Art. 25(2) can be seen as some kind of **protection against "back doors"** where controllers collect additional data, store it for longer periods, increase the extent of processing or the recipients, with the justification that it was the wish of the data subject. Evidently, data subjects who have not intervened in any way, may not even be aware of "their wishes", may not have read the expression of their wishes in detail, or are at least influenced by the default values to more likely express "wishes" favored by the controller.

[35] "In particular" indicates that the rest of the sentence is an application of the expression of the previous sentence.

This safeguard that explicitly requires the data subject's explicit intervention thus mandates the use of opt-in dialogs and prohibits opt-out dialogs. It is the same concept that is called a "clear affirmation action" in the context of consent (see Art. 4(11) GDPR). It is directly comparable to stating that without a clear affirmative action, i.e., "***without the individual's intervention****",* additional processing in terms of the amount and storage period of data, extent of processing, or number of recipients is illegitimate. It is important to note that this requirement of opt-in solutions is independent on whether *consent* is chosen as the legal basis or not.

Based on the above analysis, the measures referred to in Art. 25(2) could include the following:

- Measures that ascertain that the default settings minimize the data protection impact of the processing.
- Measures that ascertain that the data subjects are informed about the consequences of the settings that are under their control.
- Measures that ascertain that the decisions expressed by the settings are specific. For example, additional purposes cannot be enabled all with a single check-box, but it needs to be possible to enable them individually.
- Measures that verify the absence of any kind of nudging in the dialog where users chose their settings, in order to make sure that data subject can freely choose their preferences.

### 2.3.5 Applying data protection principles in the different phases of processing

The objective of data protection by design is to integrate (or implement) in all phases of a processing activity appropriate technical and organizational measures that implement data protection principles.

The EDPB's guidelines on DPbDD contain a major section on "implementing data protection principles in the processing of personal data using data protection by design and by default". It is structured by the data protection principles that have to be applied. The guidelines by the EDPB do not address the question of how to apply DPbDD in the different phases.

This section on how to apply data protection principles does not focus on a description of the principles themselves as do the EDPB guidelines (independently of phases); a detailed description of the guidelines was already provided in the according chapter of the PANELFIT guidelines (see Part II of these Guidelines, section "Principles"). In fact, this section discusses the processes that can be used to apply these principles in every of the three phases that was identified in the analysis of Art. 25(1) above.

What is thus common to all three phases is that they use the ***principles* of data *protection*** in every work step (or decision) in order to

- **identify risks** that lead to the violation or inadequate implementation of a principle, and

- **identify** appropriate technical and organizational **measures** that mitigate these risks.

The actual measures to implement largely depend on the nature, scope, context and purposes of processing. It is therefore not possible to provide a complete list of appropriate measures for each tuple of phase (or task within a phase) and principle. This section therefore describes the process of identifying appropriate measures. A detailed discussion (with examples) of measures to implement the various principles have been provided in the according section of the Guidelines.

The following discusses the phases of *determining the purposes*, *determining the means*, and the *processing itself* in more detail.

#### 2.3.5.1 Determining the Purposes

A processing activity is conceived by determining its purposes. This sets the objective of what the processing activity should achieve. This specification of "what" has to be done is still relatively abstract and lacks any detail of "how" this objective is reached. The "how" is subject to the determination of the means.

Purposes are typically determined by the top management that represents and is responsible for an organization (or organizational unit). Purposes are typically expressed in the same language in which the mission or mandate of the organization are expressed. This means, they come from the "application domain" and lack any technical content. A purpose specification falls short of determining technical decisions such as what resources (i.e., means) are needed for achieving the objectives, what data has to be collected, etc. In fact, a purpose specification can be implemented in many different ways. The objective of the determination of the means is then to find the best implementation from a data protection point of view.

According to Art. 5(1)(a) GDPR, purposes must be "specified [and] explicit". This means that they must be captured in a precise written form.

The determination of the purposes of processing is typically an iterative process. Starting with the main purpose(s), the specification is continuously completed and refined until it results in a final version. Each version has to be assessed, taking into account the data protection principles, the reasonable expectations of data subjects, and the overall risk the processing is likely to pose. Based on this assessment, improvements are made to the purpose specification which improve the observance of principles, are more balanced with the expectations of data subjects, and keep the need/benefit of the processing in balance with the risk it poses to data subjects. The iterations can be seen as a process to find the minimal impact on the rights and freedoms of data subjects while still achieving the essential objectives of the organization. Typically, in every integration, the purpose specification becomes more focused, narrower, and specific and imposes a lower impact on data subjects.

This process is visualized in the next Figure.

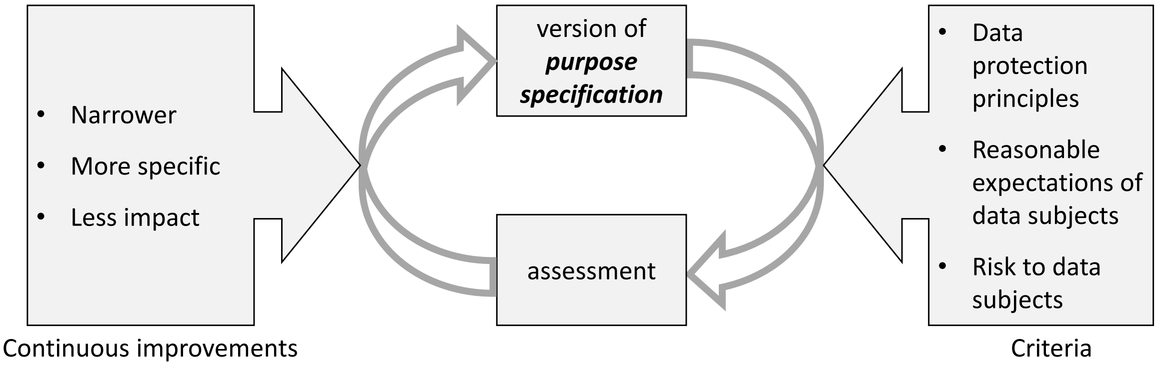


Figure 7: The process of purpose specification.

Data protection by design applies the principles of data protection to every step of determination. While some of the data protection principles are better applicable to means of processing, *legitimacy*, *lawfulness*, and *fairness* are directly applicable to purposes. Indirectly, also *data minimization* is applicable in the sense that the impact of the processing on data subjects should be minimized. This then typically results in a minimization of data that is collected about data subjects. Note also that purpose limitation during the determination of the means is only meaningful if the purposes are specified narrowly; only then it can be precisely determined whether data or processing steps are indeed necessary for the purposes. The main principles are discussed in further detail in the following.

**Lawfulness (**see “Lawfulness, fairness and transparency” in section “Principles” within Part II of these Guidelines):

According to Art. 6 GDPR, processing is lawful if one of the **legal bases** described in its paragraph 1 apply. Art. 9 GDPR adds additional requirements for special categories of data. To comply with the principle of *lawfulness*, the controller must choose a legal basis from Art. 6 and possibly 9 GDPR for every single purpose that is pursued by the processing activity.

Note that it is common that a processing activity pursues a multitude of purposes that use different legal bases. An illustration of this using the example of online shopping was described by Bruegger et. al[36].

**Legitimacy** (see “Lawfulness, fairness and transparency” in section “Principles” within Part II of these Guidelines):

While lawfulness is concerned with Art. 6 and 9 of the GDPR, legitimacy requires to follow the law in the broadest sense. It is thus not limited to the GDPR but extends to any other applicable law. Arguably, laws should not only be followed by the letter but

[36] Bud P. Bruegger, Eva Schlehahn and Harald Zwingelberg, Data Protection Aspects of Online Shopping – A Use Case, W3C Data Privacy Vocabularies and Controls Community Group, December 12, 2019, https://www.w3.org/community/dpvcg/2019/12/12/data-protection-aspects-of-online-shopping-a-use-case/ (last visited 15/7/2021).

also in spirit. In many situations, legitimacy may also be interpreted to include soft law such as commonly used ethics requirements and professional standards. It may even extend to protect the values of society at large.

The assessment of the legitimacy of purposes depends largely on the nature, scope and context of the processing. In some cases, compliance with legitimacy may require formal steps. This is for example typical in research organization where a processing activity has to be preventively approved by a research ethics committee.

**Fairness** (see "Lawfulness, fairness and transparency" in section "Principles" within Part II of these Guidelines):

A key element of fairness is to take the reasonable expectations and situations of data subjects into account. The interests of the controller, as expressed in the purpose specification, are then balanced with those of data subjects. The impact on the rights and freedoms of data subjects should be justified with an according level of necessity and potential benefits to the controller.

The assessment of the fairness of purposes typically requires assessing the expectations of data subjects. There are various ways of doing this, including just "putting oneself in the position of data subjects" up to involving consumer organizations or conducting surveys.

To assess the expectations of data subjects, it is often useful to distinguish different personae that represent different types and situations of data subjects. These should also include particularly vulnerable data subjects (such as minors or patients), or groups of data subjects who may be impacted much more significantly by the processing than the average.

The balancing must consider the risks that the processing activity represents for the rights and freedoms of data subjects. An quick overall assessment of the risk is provided by the Article 29 Data Protection Working Party's 9 criteria[37] whether a processing activity results in high risk (and therefore requires a data protection impact assessment). This should be complemented by an analysis of how special categories of data subjects and vulnerable data subjects are affected by the planned processing activity.

Note that a balancing test is formally required where the legal basis of *legitimate interest* (see Art. 6(1)(f) GDPR) was chosen for a given purpose. Guidance on how to conduct a balancing test in this context was provided by the Article 29 Data Protection Working Party[38] (see "Legitimate interest and balancing test", Part II section "Main

[37] See pages 9 – 11 in Article 29 Data Protection Working Party, WP 248rev.01, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, Adopted on 4 April 2017, As last Revised and Adopted on 4 October2017, https://ec.europa.eu/newsroom/article29/items/611236 (last visited 15/7/2021).

[38] in Article 29 Data Protection Working Party, WP217, Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC, Adopted on 9 April 2014, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf (last visited 15/7/2021).

Tools and Actions"). In a more general setting, the EDPS has provided guidelines on proportionality[39].

### 2.3.5.2 Determining the Means

The following subsection describes how to identify appropriate technical and organizational measures when determining the means.

While determining the purposes of processing specifies the "what" shall be achieved by the processing, determining the means specifies "how" this objective is achieved. In every step of determining this "how", data protection principles and requirements must be taken into account.

Determining the means can be seen to result in an *implementation plan* of the processing activity. It entails resources, instructions as well as technical and organizational measures. The latter are designed to implement the data protection principles. For a detailed discussion of measures that implement the various principles, see the Guideline section on principles guidelines (see section "Principles" within Part II of these Guidelines).

#### 2.3.5.2.1 *Managing the process of determining the means*

Determining the means is often a substantial process that typically involves a multitude of persons, fields of expertise, organizational units or departments, and may even involve external consultants and experts.

The **primary** (meta-) **organizational measure** is therefore to **set up the process of determining the means** in a way that it complies with data protection by design. This measure is referred to as a "meta-measure" since it is designed to identify those measures that actually implement the data protection principles. The meta-measure must assign clear responsibilities to the upper management:

- The upper management who represents the controller legally has to be in control of this process and mandate that data protection is appropriately taken into account in every step and decision.
- The upper management must be able to design whether the determined means (i.e., the result of this process) actually comply adequately with the data protection requirements.
- At the end of this process, it lies in the responsibility of the upper management to sign off on the determined means and give a go-ahead for the actual processing operations (the processing itself).

There are different **possible** (meta-) **organizational measures** of how to achieve this. Some examples are listed in the following:

[39] European Data Protection Supervisor, EDPS Guidelines on assessing the proportionality of measures that limit the fundamental rights to privacy and to the protection of personal data, 19 December 2019, https://edps.europa.eu/data-protection/our-work/publications/guidelines/assessing-proportionality-measures-limit_en (last visited 15/7/2021).

- Every step or decision made as part of determining the means must describe the relevant data protection requirements and how they have been enforced or otherwise satisfied.
- If a staged approach is chosen[40][41], any transition of stage gates must be subject to the approval of the data protection aspects.
- A clear designation of persons responsible for determining whether data protection requirements have been met in the individual steps should be made.
- Where available, the data protection officer[42] should be involved in the process.
- (Continuous) documentation (i.e., demonstration) of considering and incorporating data protection should be an integral part of the process. This serves both, to satisfy the principle of *accountability* (see Art. 5(2) GDPR) and as a basis for the determination by upper management for their decision to formally approve the result to be operationally used (i.e., a go-ahead for the *processing itself*).

The process of determining the means inevitably needs to **assess the effectiveness** of various measures (see discussion of effectiveness in section 2.3.3.7 above). This typically requires to perform

- risk assessments and
- surveys of the state of the art or market.

Note that the formal tool foreseen in the GDPR to assess the effectiveness of data protection measures is the *data protection impact assessment* (DPIA, see Art. 35 GDPR) (see "DPIA", Part II, section "Main Tools and Actions"). Both, risk assessment and description of measures are contained in its mandatory parts. A DPIA is only formally required by the GDPR in presence of high risk but can be used informally within the internal process. A DPIA is also a prime tool for documenting compliance with data protection by design.

At least larger organizations with several distinct processing activities can benefit from using a more **systematic approach** of determining the means. This can include the following:

- The use of ***data protection policies*** that are applicable to multiple processing activities and can thus bring economy of scale (see Art. 24(2) GDPR).
- The identification and application of applicable industry-wide codes of conduct can save effort and improve quality of implementation (see Art. 24(3) GDPR).

[40] See for example, https://en.wikipedia.org/wiki/Phase-gate_process (last visited 13/7/2021).

[41] Note that stages are not restricted to "waterfall" management, but exist also in agile methods, such as the Agile Unified Process, see http://www.ambysoft.com/unifiedprocess/aup11/html/phases.html (last visited 13/7/2021).

[42] Note that the data protection officer does not bear direct responsibility for compliance but is the internal expert likely most familiar with the requirements of the GDPR (see also Art. 39(1)(a) through (c) GDPR).

The **final result** of a successful process of determining the means is a clear and documented approval of the means and a **go-ahead** by the upper management that represents the controller. The go-ahead is necessary in order for the controller to assume full responsibility for the processing (see Art. 29 GDPR). As an additional basis for the go-ahead decision, controllers can seek **formal certification** according to Art. 42 GDPR (see Art. 24(3) GDPR). Certification represents a formal attestation of compliance with the GDPR. A documented go-ahead is a pre-requisite for the start for the operational stage of processing (the *processing itself*).

#### 2.3.5.2.2 *Assessing the effectiveness of measures relative to the data protection principles*

The above process should take a systematic approach to applying all data protection principles systematically to all decisions about means. In particular, each principle has to be enforced with technical and organizational measures. It has to be shown that these measures are effective in regard to

- "*the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing*",
- "*the cost of implementation*",
- "*the state of the art*", and
- "*the nature, scope, context and purposes of processing*"

(see section 2.3.3.7 above).

When the risk is assessed (see first bullet point), one basic risk is that the principle is violated or insufficiently guaranteed. This could be the case for all data subjects or for special groups or minorities. The vulnerable data subjects that were possibly identified during the determination of the purposes should be taken into account (see section 2.3.5.1).

To evaluate the third aspect of effectiveness, it may be necessary to conduct surveys of the state of the art.

One way to evaluate the effectiveness of measures is to use an iterative approach that is very similar to that used to determine purposes (see Figure ). Instead of a version of the *purpose specification*, a concrete *implementation plan* is evaluated. This plan entails both resources, instructions, and already foreseen technical and organizational measures (see section "Principles" in Part II of these Guidelines). In every iteration, the effectiveness of the measures is assessed and the plan is improved according to the shortcomings that were identified. The iterative process then terminates when an implementation plan with effective measures has been found.

To render this process systematic, each task that results in a decision about the means has to be evaluated in regard of all principles. Section 2.3.3.4 above has provided an overview of possible tasks. The precise breakdown of the overall determination into task depends on the nature, scope, context and purposes of the processing activity, however. It is therefore necessary to adapt the breakdown into tasks to the concrete situation.

#### 2.3.5.3 Processing itself

The following looks at applying data protection principles during the operational phase, i.e., the processing itself.

**Transparency** and **fairness** are probably the most relevant principles in this phase (see "Lawfulness, fairness and transparency" in Part II, section "Principles"). They require among others the following technical and organizational measures:

- The efficient processing of data subject right invocations.
- The handling of personal data breaches.

At the end of a processing activity, (the temporal aspect of) **data minimization** (see the "Data minimization" in Part II, section "Main Principles" of these Guidelines) requires for the personal data that is no longer necessary for the purposes to be erased. Various measures are available to ascertain that the data is irreversibly erased and that all technical storage devices are considered before their dismantling. These measures also support the principle of **purpose limitation** (see "Purpose limitation" in Part II of these Guidelines, section "Principles"): since failure to erase the data would open the possibility that they are used for other purposes. The effectiveness of the measures used for dismantling should be verified and documented as described in section 2.3.5.2.2 above.

Art. 5(1)(b) GDPR foresees the possibility of **further processing for compatible purposes**. The principle of **purpose limitation** requires careful assessment (according to Art. 6(4) GDPR) to see whether these purposes are indeed compatible. Such further processing also includes the implementation of additional measures such as further **data minimization**, pseudonymization or anonymization (i.e. **storage limitation**) in order to guarantee the **safeguards** required in Art. 89(1) GDPR.

While the **effectiveness of measures** has initially been verified during the determination of the means, the 2nd sentence of Art. 24(1) GDPR requires that this is **regularly reviewed** and that measures are updated where necessary. Such reviews and updates are measures in their own right.

Examples of where such reviews are listed in the following:

- Access rights for staff that guarantee **confidentiality** and **purpose limitation** may have to be updated to reflect staffing changes and the end of temporary assignments and substitutions.
- Software that was found to guarantee **confidentiality** may no longer do so unless critical security updates are installed.
- **Confidentiality** that was found to be sufficient may not be so anymore if the **threat landscape** evolves and **new types of attacks** become possible. Typically this requires the implementation of additional or more sophisticated measures.
- Data may have to be presumed to be **anonymous** or to prevent direct identification (as part of pseudonymization), but **new methods of re-identification** put these presumptions in question. To still support **storage**

**limitation**, a further reduction of the identification potential of the concerned data or a re-design of the processing is required.

A similar situation represents itself during the **routine replacement of** (human and technical) **resources**. When for example, a person was found to have sufficient training and skills to execute a set of instructions, the same kind of assessment is necessary for successors of this person. Similarly, new technical resources need to exhibit the same properties that guaranteed effectiveness of the original component.

Instructions typically evolve over the life time of a processing activity. Instructions for human resources and work flows may for example be re-designed or rendered more efficient based on experience. Instructions for technical resources typically change with every version of the software and often get installed automatically (e.g., by an update service). With every new version of instruction, the following has to be verified:

- That the new version still entails the measures that are necessary to guarantee effective implementation of the principles; and
- that there is no "function creep" that extends the processing beyond what is necessary for the purposes.

Where the change of resources or instructions is more substantial, a complete new iteration of the iterative process of determining the means (see section 2.3.5.2.2) may be required.

## 2.4 Identification

*Bud P. Bruegger (ULD)*

*Acknowledgements: Parts of the underlying research and writing have been conducted as part of the Forum Privatheit project which has received funding from the German Federal Ministry of Education and Research (BMBF). Discussions with and feedback on an extended version from Marit Hansen, Harald Zwingelberg and Benjamin Bremert, all ULD, and Jessica Schroers, KU Leuven, were very helpful and helped improve the content. Jessica Schroers also reviewed this shortened version of the document.*

*The final version of this section was validated by Hans Graux, guest lecturer on ICT and privacy protection law at the Tilburg Institute for Law, Technology, and Society (TILT) and at the AP Hogeschool Antwerpen. President of the Vlaamse Toezichtscommissie (Flemish Supervisory Committee), which supervises data protection compliance within Flemish public sector bodies.*

When discussing identification, a certain vocabulary is necessary. To support precise statements, this subsection begins with some definitions of terms. It then presents a technical model for the GDPR-concept of identification. Finally, it presents transformations suited to reduce the identification-potential of personal data. The latter are used in both, pseudonymization and anonymization. They are often called "anonymization techniques"; but this is misleading, since they frequently fail to guarantee that the resulting data are indeed anonymous.

### 2.4.1 Some underlying concepts

The following establishes a vocabulary of precisely defined terms that enable to make precise statements about identification, pseudonymization, and anonymization.

This subsection is designed to be read at different levels of detail. In its minimal use, it can be totally skipped and used only as a glossary when the need arises to better understand terms used in later sections. Instead of reading the full text, it is possible to abbreviate the reading by considering only the definition boxes. For brevity, this short version avoids to incorporate the discussion of how the concepts relate to the GDPR; readers interested in that aspect are referred to the more detailed analysis (see https://uldsh.de/PseudoAnon).

The discussion of pseudonymization and anonymization focuses on (personal) information.

Definition: ***information***

Information consists of expressions of facts 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.

Of particular interest is individual-level information.

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*. Since individual-level information relates to a person, it is personal information.

Individual-level information is closely related to a data record.

Definition: ***data record***

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

Personal information is composed of information elements.

Definition: ***information element***

Information elements are components of a single data record. They can either be single data values (such as the "age") or tuples of values (such as a *postal address*) that can be broken up further into smaller data elements (such as *street name*, *street number*, *postal code*, and *town*.

The following types of information elements often take on different roles in the context of pseudonymization and anonymization:

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 (such as a person) 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 establish a term to denote this context:

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, for example, is 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.

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.

Transformations to reduce the identification potential of data (often misleadingly called "anonymization techniques") often assign a special role to 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.

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 Art. 29 Data Protection Working Party in their *Opinion on Anonymization Techniques*[43] but without a clear definition.

Typical examples for quasi-identifiers are the following:

- Name, gender, date and place of birth[44];
- 5-digit ZIP, gender, and date of birth[45];
- Mobility data[46];
- 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[47].

[43] Article 29 Data Protection Working Party, WP 216, Opinion 05/2014 on Anonymization Techniques, Adopted on 10 April 2014, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (last visited 24/06/2021).

[44] This combination is for example used in some national unique schemes for unique handles such as the Italian tax number.

[45] 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).

[46] 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

The third major type of information element (after *unique handle* and *quasi-identifier*) 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 "one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"[48].

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[49] and is so rare[50] that it could easily identify a single individual.

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[51].

[47] 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.

[48] Art. 4 (1) GDPR .

[49] 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).

[50] 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, last visited 10/11/2020).

[51] 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 (last visited 10/11/2020).

Uniqueness seems to be very common in so-called *high-dimensional data sets*[52].

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 one of the 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 its closest neighbors (i.e., data records represented as points). 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[53] 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.

Also, the Article 29 Data Protection Working Party emphasizes the identification potential of high-dimensional data in their *Opinion on Anonymization Techniques*[54]. 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[55] that was linked against the Internet Movie Database.

The concept of identification is closely related to that of linking.

Definition: ***linking***

[52] 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).

[53] 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).

[54] See page 30 in, WP216, footnote 43.

[55] 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 (last visited 15/12/2020).

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.

What kind of linking is possible depends on the kind of attribute value. Therefore we distinguish two kinds of values:

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[56]. 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; continuous values can be similar, close, or correlated.

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 this Figure:

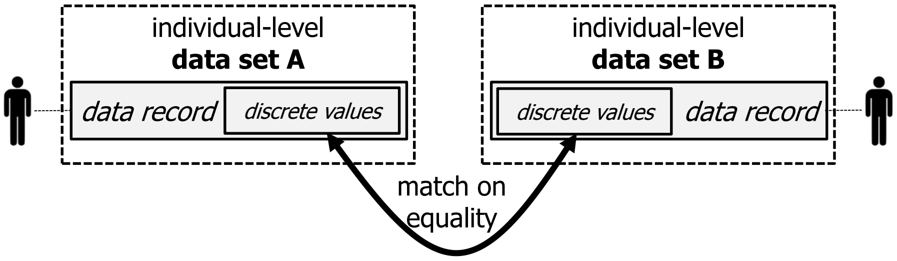


Figure 8: Deterministic linking.

[56] The difference is usually defined in terms of a distance function.

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 9:

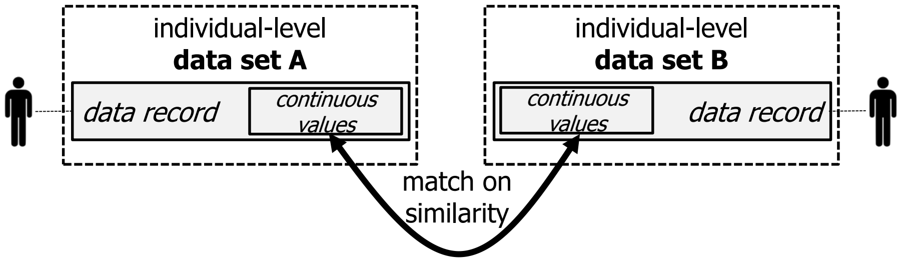


Figure 9: Probabilistic linking.

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"). In fact, the relation expresses a probability that the data records actually belong to the same person.

Note that linking by comparison of the same attribute values in two data sets is only the most common case. There are other linking methods, as for example models that establish the degree of correlation between different kinds of attributes in the data sets. Such models could for example be created though machine learning.

### 2.4.2 Definition of Identification

(Direct and indirect) identification takes a central role in the definition and understanding of pseudonymization and anonymization. It thus requires a precise analysis.

For this purpose, the following provides a technical interpretation of what happens when a person is identified in a data set. This is done in terms of actions that successful identification enables, rather than in terms of which data elements are necessary to achieve direct identification.

This approach attempts to be more precise and general than most texts on the argument, including the Article 29 Data Protection Working Party's *Opinion 4/2007 on the concept of personal data*[57]. The latter states that "[…] in practice, the notion of 'identified person' implies most often a reference to the person's name."[58]

Defining the meaning of successful identification of a person (equivalent to direct identification) in terms of elements contained in the data set, often the *name*, represents the probably most common case. This approach fails to explain why the *name* leads to identification, however. Nor does it answer the question of exactly which other data elements can also lead to identification, under what circumstances, and why.

In an attempt to understand the concept more deeply, the model proposes actions that become available to an actor only if a person has been successfully identified in the data. This model can explain why the *name*, in many common circumstances, leads to identification. Beyond just the *name*, the model is also applicable to other data elements.

The model of identification is illustrated in Figure  that is described in the sequel.

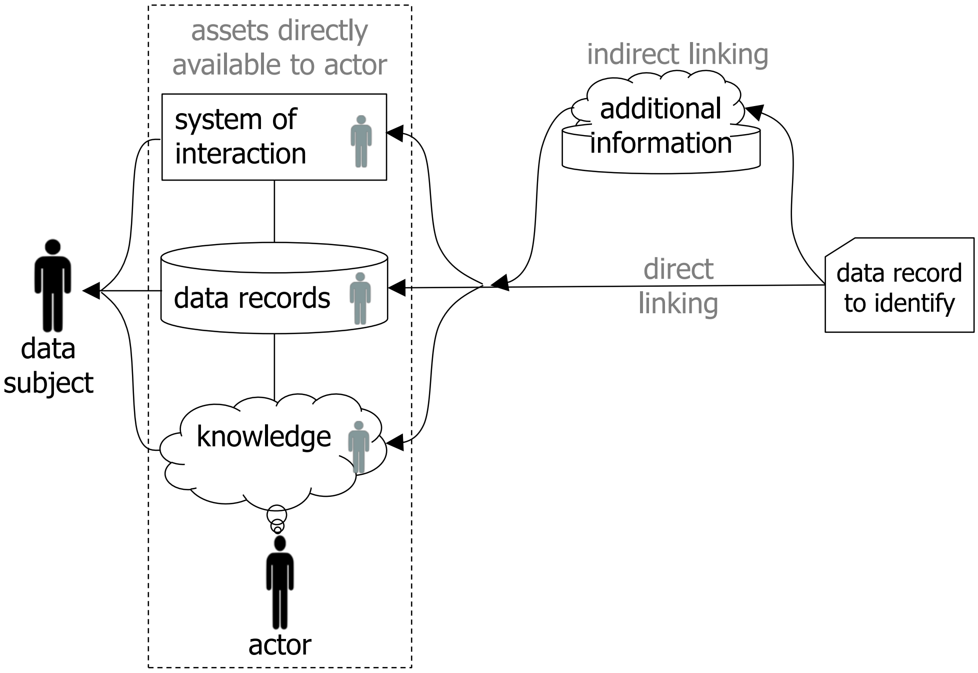


Figure 10: Identification of a data subject.

[57] Article 29 Data Protection Working Party, WP136, Opinion 4/2007 on the concept of personal data, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf (last visited 28/6/2021).

[58] Idem, 2nd paragraph on page 13.

Identification is about relating a **record of a data** set, shown on the right, with a **data subject**[59], shown on the left. Identification requires an **actor** who attempts the identification. Actors have certain **assets** at their direct disposition relative to which the identification is performed.

These assets include **information assets** consisting of

- knowledge of data subjects in the mind of the actor, and
- data records representing data subjects in some processing system.

These information assets can be seen as a **virtual model of the world** that includes representation of real persons.

In addition to information assets, actors also have access to **systems that permit to interact with persons**. The most common examples of such system may be communication systems such as telephone, e-mail, messaging, or postal mail. Actors can also interact physically with persons by meeting with them.

Actors have successfully identified a data subject in the data record when they are enabled to perform certain actions, namely the following:

- Actors are able to consult and/or manipulate the representation of the data subject in their virtual model of the world represented by data or knowledge contained in their information assets.
- Actors are able to interact with the physical person through a system of interaction that is available to them.

The former kinds of actions are for example enabled when the matching virtual representation of the data subject in the information assets can be established through lookup[60] (e.g., based on a unique handle contained in the data record) or recognition[61] (e.g., based on a unique combination of identity-relevant properties). It is evident that a *name* is in many cases a suitable handle to look up persons in information assets. It is also clear that this is only the case when the name is contained in the information assets. Furthermore, information elements different from a name can enable lookup or recognition.

The latter kinds of actions are typically enabled when the data record contains an address that identifies a data subject in a given system of interaction. Addresses are in most cases unique handles in the identity domain defined by the system of interaction. It can also be a time and place, however, that permits to meet and physically interact with a person. In some cases, it may be necessary to add additional information elements to

[59] Note that in the general case, instead of an individual data subject, the relation could also be made to a class of data subjects or to a session. The model is thus also applicable to c- and s-Identification proposed by Leenes in: R. Leenes, 'Do They Know Me? Deconstructing Identifiability' (2008) 4(1&2) University of Ottawa Law &
Technology Journal 135, 141-142, https://pure.uvt.nl/ws/portalfiles/portal/1310856/Leenes_Do_they_know_me_110216_publishers_immediately.pdf (last visited 29/6/2021).

[60] This corresponds to l-identification proposed by Leenes.

[61] This corresponds to r-identification proposed by Leenes.

an imprecise time and place that allow for the recognition of the person with whom to interact. For example, a description or picture of the person may serve this purpose. Other examples are unique properties like a description of what a person wears or carries[62].

The relevant concepts of identification are formalized in the following definitions.

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) in the mind of a human actor,
- used in a virtual model of the world (i.e., data) available to the actor, or
- used 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 (unique) combinations thereof. It results in the association between the data record and a mental representation, data record, or interaction address of the related person in the domain of the actor.

Identification is considered to be **direct** if it happens solely based on the assets that are directly available to an actor. These include the knowledge and data the actor possesses, as well as other assets that are at ready disposition such as those resulting from a simple internet search or phone book lookup.

Definition: ***directly available assets***

An asset is considered directly available to an actor if the actor knows about its existence and can access it with contained effort. Most prominently, this is the case for assets that are under the direct control of the actor.

Definition: ***direct identification***

Direct identification is based on linking between the data record and a unique handle contained in directly available assets.

[62] Such as “a red carnation in the buttonhole a copy of the times under the left arm”.

Identification is considered to be **indirect** if it is only possible with assets that the actor cannot readily access. Such assets are typically called *additional information*. Similarly, information is considered additional if it requires a significant effort on part of the actor to gain access to the information.

Definition: ***not directly available assets; additional information***

An asset is considered not directly available to an actor if the actor initially does not know about its existence or can access it only with significant effort. Not directly available information assets are typically called *additional information*.

Definition: ***indirect identification***

Indirect identification is based on multi-step linking between the data record via not directly available assets to a unique handle contained in directly available assets. In most cases, the initial data record is first linked to additional information and from there to directly available assets.

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[63].

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.

### 2.4.3 Reducing the identification potential of data

There are many factors that influence how easily data subjects can be identified in a given data set. For example, whether potentially motivated actors have access to a data set or whether it is protected by confidentiality strongly influences the potential of identification. In the context of pseudonymization and anonymization, another factor is highly prominent, namely the identification potential of the data themselves.

While it is difficult to quantify the identification potential of data, the present section gives an overview of transformations that yield a new data set with reduced identification potential. The more detailed version of this analysis[64] provides further detail. Alternative surveys of methods are for example provided by S. Garfinkel (published by NIST)[65] and Fung et al.[66].

[63] This definition aims at being in line with the sentence 3 and 4 of Recital 26 GDPR.

[64] See https://uldsh.de/PseudoAnon.

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

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.

In all the described scenarios, the linking takes place between two sets of data (or information). One of these is an asset directly available to 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.

#### 2.4.3.1 Prevention of 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.

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

(i) Deletion of all unique handles from the data set;

(ii) Replacement of unique handles with unique handles (aka pseudonyms) 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 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.

[66] 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 (last visited 23/12/2020).

A more detailed discussion of options can be found in the ENISA report on pseudonymization[67].

#### 2.4.3.2 Prevention of 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 a quasi-identifier such that the remainder of the quasi-identifier becomes less unique;
- **Generalization** of the values that the quasi-identifier is composed of.

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[68] 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*[69]. The latter illustrates that the intervals do not need to be regular.
- Ordinal-scale[70] 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****.

[67] See European Union Agency for Cybersecurity (ENISA); Athena Bourka, Prokopios Drogkaris, and Ioannis Agrafiotis (all ENISA, Editors); Meiko Jensen (Kiel University), Cedric Lauradoux (INRIA), Konstantinos Limniotis (HDPA) (contributors); Pseudonymization techniques and best practices; Recommendations on shaping technology according to data protection and privacy provisions; November 2019; https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices (last visited 12/8/2021).

[68] See for example https://en.wikipedia.org/wiki/Level_of_measurement#Ratio_scale (last visited 25/11/2020).

[69] See for example https://en.wikipedia.org/wiki/Generation#Western_world (last visited 25/11/2020).

[70] See for example https://en.wikipedia.org/wiki/Level_of_measurement#Ordinal_scale (last visited 25/11/2020).

- Nominal-scale[71] 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 11 that was taken from Kristen LeFevre et al.[72].

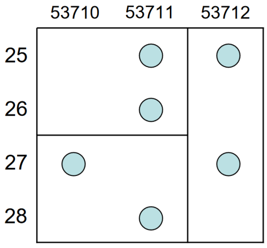


Figure 11: Example of a two-dimensional generalization with ZIP code and age by LeFevre et al.

The most common method to assess whether deletion and generalization in quasi-identifiers sufficiently impedes linkability is ***k-anonymity***[73] 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*[74].

[71] See for example https://en.wikipedia.org/wiki/Level_of_measurement#Nominal_level (last visited 25/11/2020).

[72] 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).

[73] See Samarati, P. and L. Sweeney, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, 1998, https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf (last visited 12/8/2021).

[74] See for example 2nd 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, (last visited 22/12/2020).

#### 2.4.3.3 Prevention of 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 on 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.

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.

##### 2.4.3.3.1 *Truthful concepts of transformation*

The following describes truthful concepts of transformation.

**Deletion**

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

Deletion can affect different data elements:

- A **single attribute** belonging to a **single data subject** when a value of an attribute becomes too rare (e.g., a very high age).
- A **single attribute** across **all data subjects**, i.e. in tabular data, this would delete a whole **column**.
- **All attributes** belonging to a **single data subject**, for example 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** (aka *cell suppression*), for example when for a given cell k (of k-anonymity) cannot be reached.
- **Resampling** of time series.

### Generalization

Generalization was already discussed above for quasi-identifiers. The same transformation concept can be applied also to identity-relevant properties.

The following distinguishes different kinds of generalization:

- Generalizations that apply **coarser scales of measurement** to **one or several attributes** in individual-level data, for example:
    - Rounding a precise continuous value to a lower precision.
    - Aggregating sets of nominal-scaled values into categories.
    - Grouping point locations (e.g., latitude and longitude) into areas (such as ZIP code areas, census districts, provinces, or countries).
- Generalizations that map **multiple attributes of a single data subject** to coarser **statistical attributes**, for example:
    - A time series of a patient's body temperature mapped to the average, minimum, and maximum temperature.
- Generalizations that map attributes of **multiple data subject** to a **single attribute describing groups of data subjects:**
    - Statistics.

Note that it may be common to think that it was impossible to link statistical data to individual-level data sets; in other words, that statistical data were free of risk of identification. As is described well by Garfinkel et al.[75], 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.

### Slicing

It has been argued above that multi-dimensional data has a higher 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.

[75] See footnote 74.

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 is then 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, very similar to k-anonymity.

Examples of slicing include the following:

- ***Anatomization*** proposed by Xiao and Tao[76].
- **KC-slice method** by Onashoga et al.[77] that was further refined by Raju et al.[78].

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.

#### 2.4.3.3.2 *Concepts of transformation that introduce deviations from the truth*

This section provides an overview of transformations that introduce deviations from the truth.

**Top- and bottom-coding**

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. Top- and bottom-coding are routinely used in statistical publications such as census data[79].

**Data swapping**

The basic concept of *data swapping* 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 "group" or "cell". For a more detailed discussion of data

[76] 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).

[77] 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,

[78] 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).

[79] See page 3 of Simson L. Garfinkel, De-Identifying Government Datasets, NIST Special Publication 800-188 (2nd Draft), 2016, https://csrc.nist.gov/publications/detail/sp/800-188/draft (last visited 6/1/2021).

swapping, see for example Fienberg and McIntyre[80]. A description how data swapping was used in the U.S. 1990 census is provided by McKenna[81].

### Random noise injection

In *noise injection* (aka *noise addition*), a random error is added to truthful data. The more error is added, the less likely that identification is still possible

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 proposed by Dwork et al.[82]. According to Sartor[83], "[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"[84]. Referencing Dwork et. al[85], Wikipedia states: "Although it does not directly refer to identification and reidentification attacks, differentially private algorithms probably resist such attacks."[86] This statement seems to be further supported by McClure and Reiter[87].

Sartor provides a good introduction to the topic and lists further introductory resources; a more detailed introduction was provided by Wood et al[88].

Differential privacy is not a single transformation to reduce the identification potential of a data set. In fact, differential privacy is a mathematical framework that is based on a mathematical definition of what privacy actually is. There are hundreds of published

[80] Fienberg, S. and J. McIntyre. "Data Swapping: Variations on a Theme by Dalenius and Reiss." Privacy in Statistical Databases (2004), https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/data-swapping-variations-on-a-theme-by-dalenius-and-reiss.pdf (last visited 12/8/2021).

[81] 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).

[82] 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.

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

[84] 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).

[85] See footnote 82.

[86] https://en.wikipedia.org/wiki/Differential_privacy (last visited 13/8/2021).

[87] 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).

[88] 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).

differentially private mechanisms for which there are mathematical proofs that comply with the mathematical framework. Examples include building a histogram[89], taking an average[90], releasing micro-data[91] (i.e., individual-level data), and generating a machine learning model[92].

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[93]. Among the most prominent practical applications of differential privacy is the Census 2020[94] by the U.S. Bureau of the Census and the mining of user data by Apple[95] [96]. But neither of these have yet proven to be success stories.

Sartor concludes his experience by saying that differential privacy was "beautiful in theory" but a "fallacy in practice"[97]. 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."

#### 2.4.3.4 Key aspects 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

[89] 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://web.cs.ucdavis.edu/~franklin/ecs289/2010/dwork_2008.pdf (last visited 13/1/2021).

[90] 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).

[91] 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).

[92] 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).

[93] 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).

[94] 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_confi.html (last visited 15/1/2021).

[95] WWDC 2016. June, 2016.Engineering Privacy for Your Users. https://developer.apple.com/videos/play/wwdc2016/709/. (last visited 15/1/2021).

[96] WWDC 2016. June, 2016.WWDC 2016 Keynote.https://www.apple.com/apple-events/june-2016/. (last visited 15/1/2021)

[97] See footnote 83.

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 have 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.
- 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.
- Although these transformations are often referred to as ***anonymization techniques***, they **fail to guarantee** that the result of the transformation is **indeed *anonymous***.

#### 2.4.3.5 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 relevant 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)[98],
- Aircloak[99],
- Johns Hopkins University[100],
- YourTechDiet[101], and

[98] https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/focus-areas/de-id/tools (last visited 21/1/2021).

[99] https://aircloak.com/top-5-free-data-anonymization-tools/ (last visited 21/1/2021).

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

[101] https://www.yourtechdiet.com/blogs/6-best-data-anonymization-tools/ (last visited 21/1/2021).

- Electronic Health Information Laboratory (EHIL)[102],

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, ℓ-Diversity, t-Closeness, and differential privacy[103].

## 2.5 Pseudonymization

*Bud P. Bruegger (ULD)*

*Acknowledgements: Parts of the underlying research and writing have been conducted as part of the Forum Privatheit project which has received funding from the German Federal Ministry of Education and Research (BMBF). Discussions with and feedback on an extended version from Marit Hansen, Harald Zwingelberg and Benjamin Bremert, all ULD, and Jessica Schroers, KU Leuven, were very helpful and helped improve the content. Jessica Schroers also reviewed this shortened version of the document.*

*The final version of this section was validated by Hans Graux, guest lecturer on ICT and privacy protection law at the Tilburg Institute for Law, Technology, and Society (TILT) and at the AP Hogeschool Antwerpen. President of the Vlaamse Toezichtscommissie (Flemish Supervisory Committee), which supervises data protection compliance within Flemish public sector bodies.*

This section describes the GDPR's concept of *pseudonymization* and how to implement it. Pseudonymization is a manner of processing of personal data. Pseudonymization is concerned with rendering identification in a controlled environment of a controller's (or joint controllers') processing activity (or activities) (or subset thereof) impossible. This requires (among others) that the data are rendered pseudonymous in a manner that they do no longer permit direct identification of data subjects.

*Pseudonymization* contrasts with *anonymization* which renders both direct and indirect identification impossible in all possible environments.

In order to understand the concept of pseudonymization, the following analysis attempts to provide a detailed conceptual framework with a precise technical interpretation. For this purpose, it defines precise (technical) meanings for the terms used in the GDPR and where necessary, introduces additional concepts and distinctions. The terminology is

[102] http://www.ehealthinformation.ca/faq/de-identification-software-tools/ (last visited 21/1/2021).

[103] https://arx.deidentifier.org/overview/privacy-criteria/ (last visited 12/8/2021).

aimed to be compatible with the GDPR, however; re-definition of terms with a different meaning from that of the GDPR have been avoided.

***Pseudonymization* in a nutshell:**
Considering that (identified) personal data can be seen as consisting of both, a “who” and a “what” part, *pseudonymization* is a **manner of processing** that strictly separates the “who” and the “what” part such that

- the processing is limited to the “what” part and
- the “who” part is separated and protected such that it cannot be used for the identification of data subjects.

The separation of “who” and “what” based on identified data is achieved by *data pseudonymization*. *Data pseudonymization* is a transformation[104] of data. It is distinct from *pseudonymization* (as defined in the GDPR) which is a manner of processing that acts on (already) *pseudonymized* data.

In the realm of *pseudonymization*, any identification is prohibited; while the possibility of re-identification is explicitly foreseen in the GDPR, rendering the data identified again in this way exits the realm of pseudonymization and enters that of processing identified data.

The risk inherent in identified data is usually higher than the sum of the risks inherent in pseudonymized data (“what”) and the additional information (“who”). This is evident when considering that that the separation either informs “who” is in the data set or that an unknown entity has certain properties (“what”).

### 2.5.1 Motivation to use pseudonymization

There are three main reasons to use pseudonymization:

- It is required by the GDPR wherever the purposes permit it,
- it substantially reduces the risk for data subjects, and
- it therefore permits controllers to reduce the effort of implementing (other) technical and organizational measures.

### 2.5.2 Definition of 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 that visualized in Figure 12:

[104] This transformation also constitutes processing according to Art. 4(2) GDPR but is not part of the processing that constitutes *pseudonymization* according to Art 4(5) GDPR. Note that the literature often does not make the distinction between *data pseudonymization* and *pseudonymization*; the present document makes this distinction explicit for conceptual clarity.

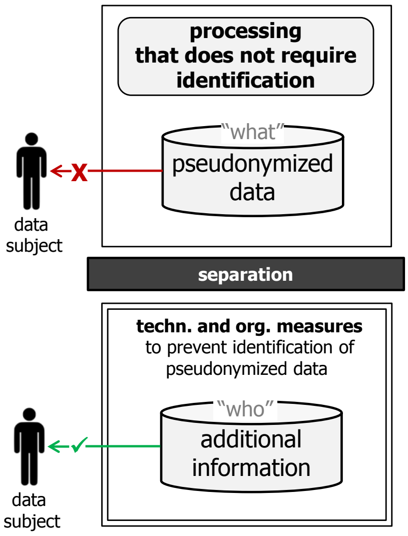

processing that does not require identification
"what"
pseudonymized data
data subject
separation
techn. and org. measures to prevent identification of pseudonymized data
"who"
additional information
data subject

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 organizational 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 (beyond storage) 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 2.5.6 below.

### 2.5.3 The context of pseudonymization

Art. 4(5) chose to define *pseudonymization* in a very narrow manner. It is however useful to see it in its wider context which includes the processing of identified data that can precede pseudonymization and a possible re-identification of data that can occur after or in parallel to pseudonymization. For this purpose, Figure  illustrates the situation.

In the middle of the Figure, the representation of *pseudonymization* from 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 below. 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*.

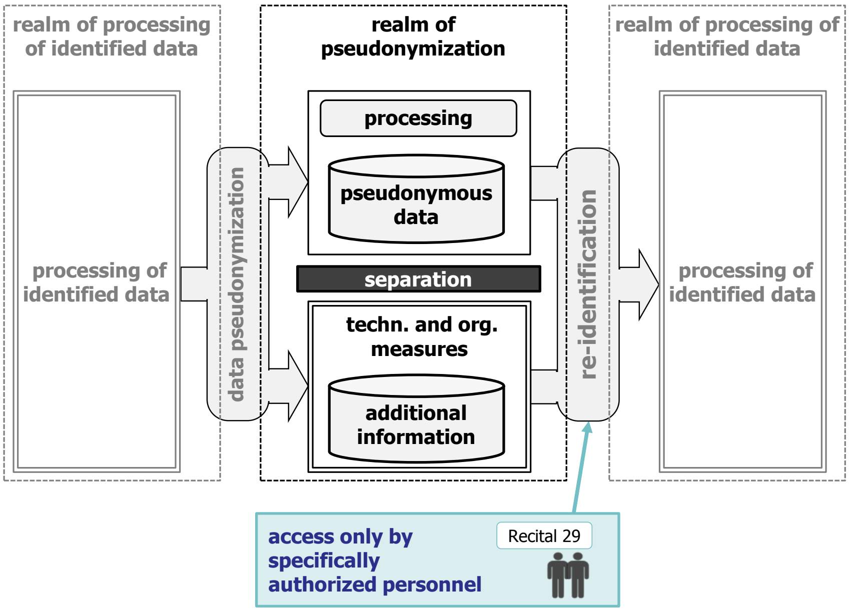

realm of processing of identified data
processing of identified data
data pseudonymization
realm of pseudonymization
processing
pseudonymous data
separation
techn. and org. measures
additional information
re-identification
realm of processing of identified data
processing of identified data
access only by specifically authorized personnel
Recital 29

Figure 13: The context of pseudonymization.

### 2.5.4 Some concepts relevant to pseudonymization

To be able to make precise statements about pseudonymization, this subsection defines some important concepts (terms).

This subsection is designed to be read at different levels of detail. In its minimal use, it can be totally skipped and used only as a glossary when the need arises to better understand terms used in later sections. Instead of reading the full text, it is possible to abbreviate the reading by considering only the definition boxes. For brevity, this short version avoids to incorporate the discussion of how the concepts relate to the GDPR; readers interested in that aspect are referred to the more detailed analysis (see https://uldsh.de/PseudoAnon).

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 .

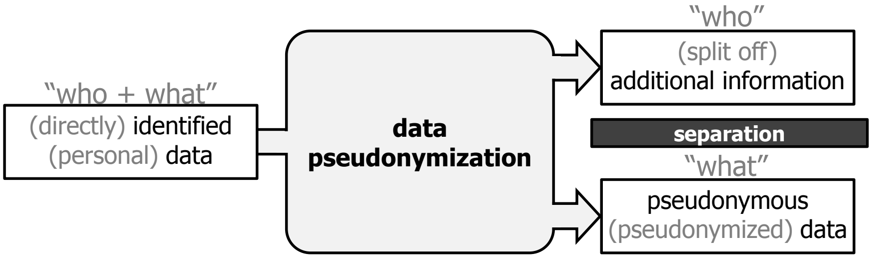


Figure 14: Data pseudonymization.

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. In fact, **any** *additional information* can be used, including and most commonly that existing outside of the controller.

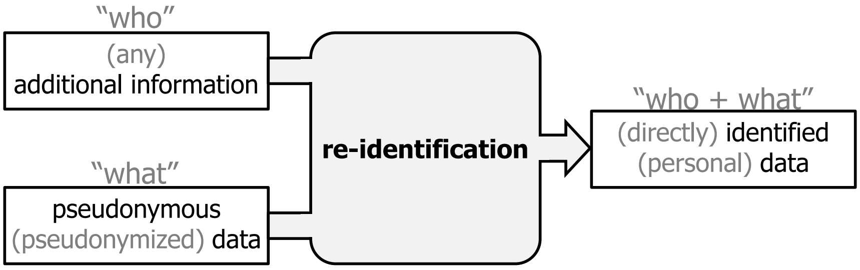


Figure 15: General re-identification.

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.

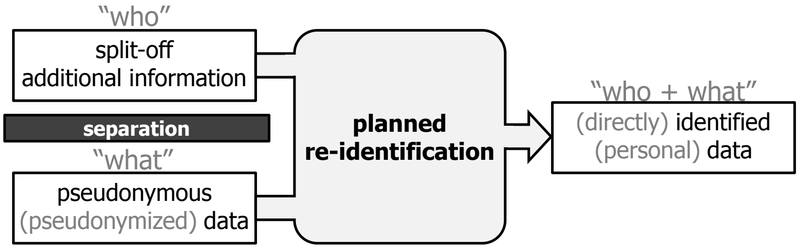


Figure 1: Planned re-identification.

Definition: (***Directly) identified***[105] ***(personal) data***

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

This is the case, for example, 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 section on identification above).

The concept of identified personal data is illustrated in Figure.

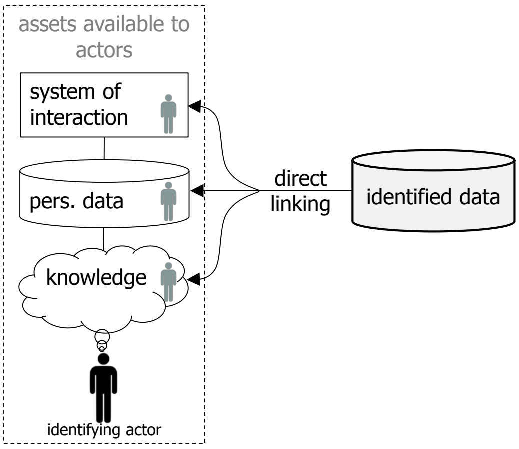


Figure 17: Identified data.

Within the concept of pseudonymous data, two types are distinguished:

Definition: ***(General) pseudonymous data*** [captures the common use of the term]

*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.

Definition: ***Strictly pseudonymous data*** [captures the use of the term in the GDPR]

*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

[105] 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.

*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*.

The concept of strictly pseudonymous data is further illustrated in Figure 18.

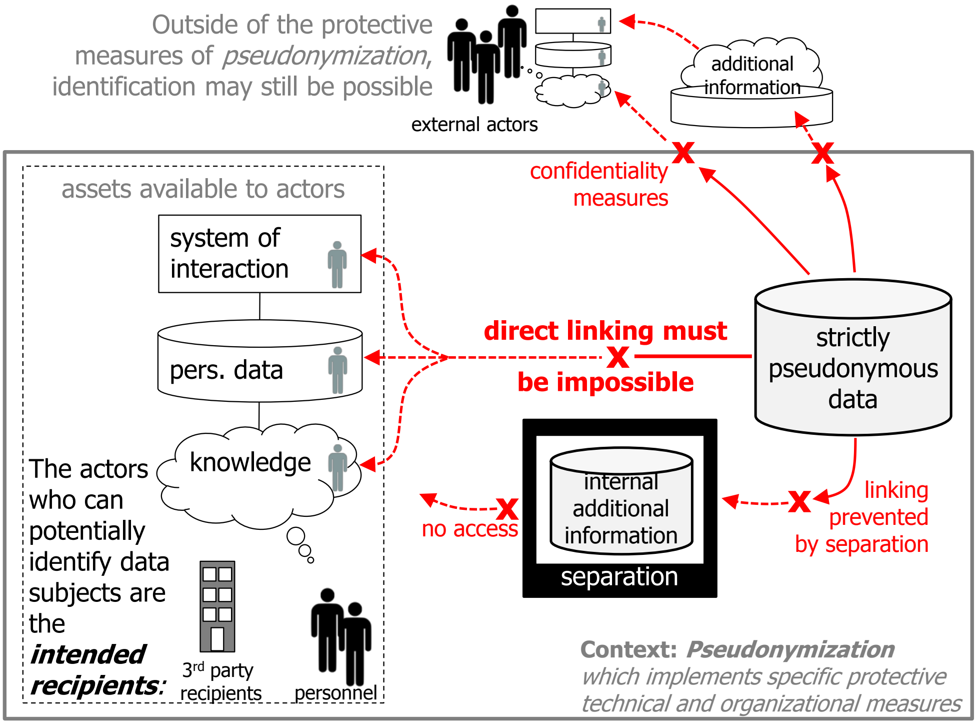

Outside of the protective measures of *pseudonymization*, identification may still be possible
external actors
additional information
confidentiality measures
assets available to actors
system of interaction
pers. data
knowledge
direct linking must be impossible
strictly pseudonymous data
The actors who can potentially identify data subjects are the ***intended recipients***:
3rd party recipients
personnel
internal additional information
separation
no access
linking prevented by separation
**Context:** ***Pseudonymization*** *which implements specific protective technical and organizational measures*

Figure 18: In the context of pseudonymization, intended recipients are unable to identify data subjects in strictly pseudonymous data.

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.

The concept of *additional information* was already defined in the context of (indirect) identification above and is further specialized here in the context of pseudonymization. Two types of additional information are distinguished based on their relationship to pseudonymization:

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 19illustrates 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.

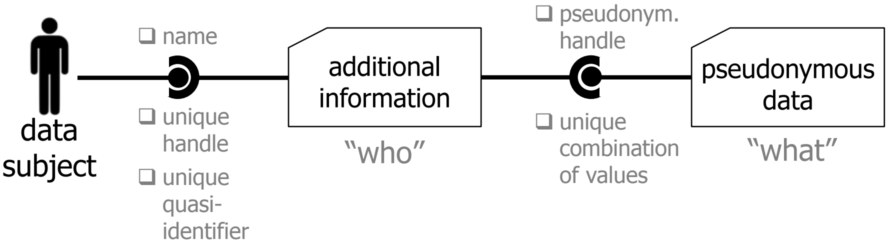


Figure 19: Additional information links pseudonymous data to a data subject.

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.

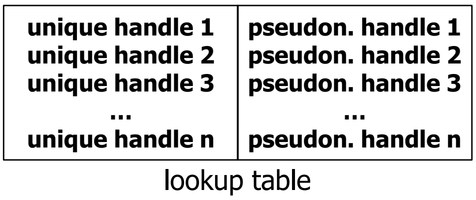
While *additional information* in general identifies at least one data subject in the set of *pseudonymous data*, *split-off additional information* usually identifies all data subjects in the set of *strictly pseudonymous data*.

While the above distinction of types of *additional information* was made based on the relationship of this information to the pseudonymization, types can also be distinguished based on the format of the information:

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 20 gives an example for *lookup-based additional information*.



| | |
|---|---|
| **unique handle 1** | **pseudon. handle 1** |
| **unique handle 2** | **pseudon. handle 2** |
| **unique handle 3** | **pseudon. handle 3** |
| **...** | **...** |
| **unique handle n** | **pseudon. handle n** |

lookup table

Figure 20: 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.

Figure 21 illustrates an example of *formula-based additional information*.

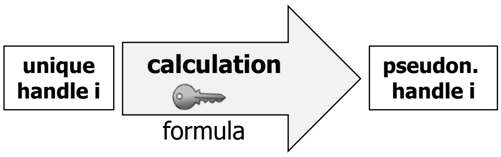


Figure 21: Formula-based additional information.

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*.

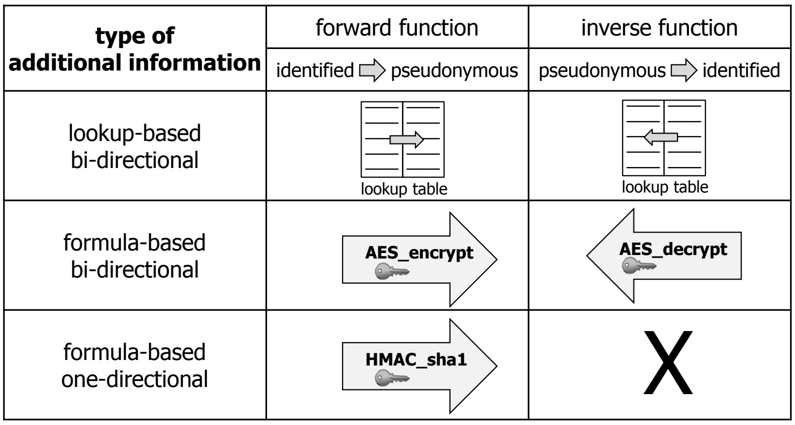
Definition: ***One-directional additional information***

*One-directional additional information* permits the use of 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  further illustrates the different types of *additional information* by providing common examples.



| **type of additional information** | forward function | inverse function |
|---|---|---|
| | identified ⇨ pseudonymous | pseudonymous ⇨ identified |
| lookup-based bi-directional | lookup table | lookup table |
| formula-based bi-directional | **AES_encrypt** | **AES_decrypt** |
| formula-based one-directional | **HMAC_sha1** | X |

Figure 22: Examples of different types of additional information.

Also for the concept of pseudonyms, two types can be distinguished:

Definition: ***(General) pseudonym*** [captures the common use of the term]

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

Definition: ***Pseudonymous handle*** [captures the meaning in the context of pseudonymization]

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*.

### 2.5.5 Data pseudonymization in detail

The following describes a typical procedure of how to perform *data pseudonymization*. In other words, it describes the steps 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 23 and discussed in the following.

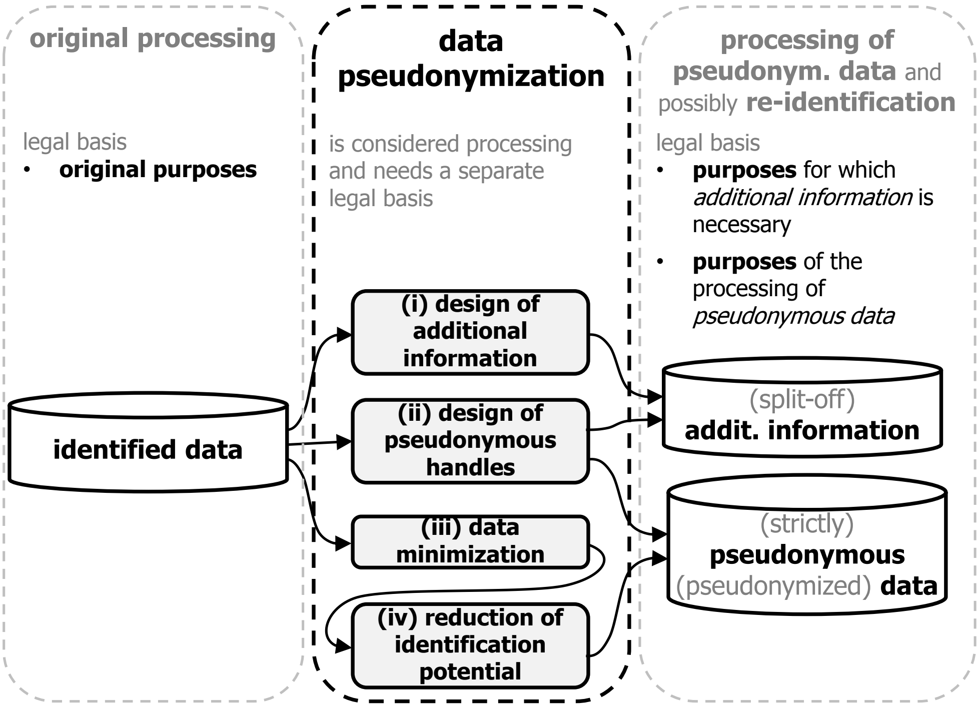

original processing
legal basis
original purposes
identified data
data pseudonymization
is considered processing and needs a separate legal basis
(i) design of additional information
(ii) design of pseudonymous handles
(iii) data minimization
(iv) reduction of identification potential
processing of pseudonym. data and possibly re-identification
legal basis
purposes for which additional information is necessary
purposes of the processing of pseudonymous data
(split-off) addit. information
(strictly) pseudonymous (pseudonymized) data

Figure23: Functional details of data pseudonymization.

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

- 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*[106] step (iii), since it filters out all data and detail that is unnecessary to fulfill the stated new purposes.

It is similarly crucial to the step of *design of additional information* (i). More precisely, 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.

[106] Note that data minimization is one of the principles of data protection (see Art. 5(1)(c) GDPR).

(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.

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 kept. 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* 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 be the case, for example, 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).

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. 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 2.4.3.1 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 though generalization. An example for the latter is to generalize a precise location (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 2.4.3 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[107].
- **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 2.4.3.4, thee 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*.

[107] 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[108] available to them. These information assets can include the following:

- **Other data** kept by the controller for other processing activities that is also accessible[109] 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**[110] 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[111]. Since only the linking methodology that is reasonably likely used[112] by the known actors (i.e., intended recipients) has to be considered, complex linking methods can often be excluded. Organizational measures that prohibit[113] personnel to attempt any linking may further exclude possibilities of identification.

### 2.5.6 Technical and organizational measures for pseudonymization

The following provides more detail on additional 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

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] See Recital 26, sentence 3, GDPR.

[113] This can for example be achieved through a contractual agreement and reinforced through training.

*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).
- 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***[114] provides an even stronger separation. These entities can then better defend the interests of data subjects, potentially even against the interests of the controller.

[114] 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.

- 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 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 including authentication, authorization and maybe logging of access[115].
- 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 employees are working with pseudonymous data about a group of persons to which they belong 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

[115] 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.

**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.
- The controller should consider to specify a **procedure to handle the case where an employee recognizes** (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[116]. Further, it may be considered to notify the concerned data subject of the "breach"[117].
- **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 all-together.
- 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 be achieved, for example, by **training** or through a **contractual obligation** with the employees.

[116] 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.

[117] 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.

- To separate the pseudonymous data from **additional information[118] 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[119] on the computing systems where pseudonymous data reside.

### 2.5.7 Different types of (re-) identification

There are different kinds of (re-)identifying data subjects in *pseudonymous data*. The various possibilities are illustrated in Figure 24. They are described in more detail in the extended version of this analysis (see https://uldsh.de/PseudoAnon).

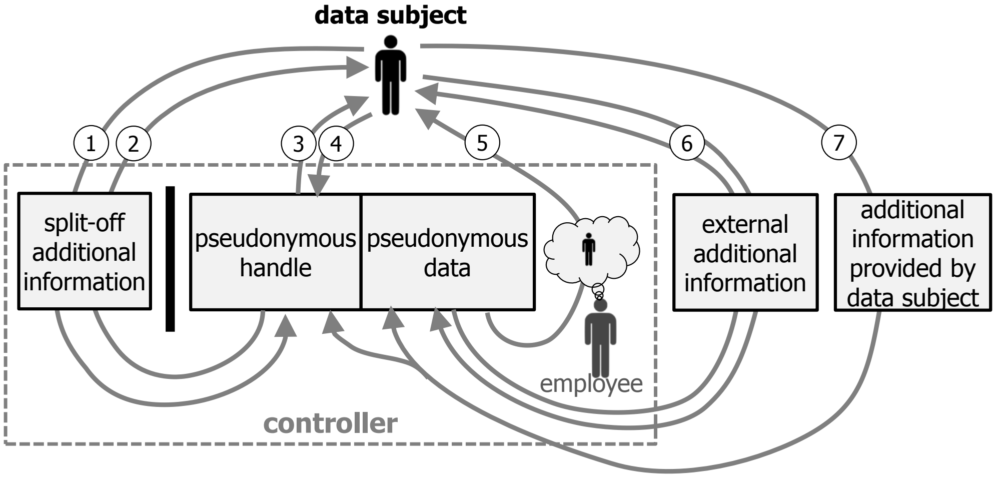

data subject
1
2
3
4
5
6
7
split-off additional information
pseudonymous handle
pseudonymous data
external additional information
additional information provided by data subject
employee
controller

Figure 24 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.

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

(3) Pseudonym inversion attack.

(4) Pseudonym creation attack for known data subjects.

[118] Note that this is different from the split-off additional information that is created as an output of data pseudonymization.

[119] Note that so called “portable” software does not require installation but can be directly used for example from a USB stick.

(5) Unexpected recognition of data subject by personnel.

(6) Indirect identification attack though the linking with external additional information.

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

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.

### 2.5.8 Pseudonymization and Art. 11 GDPR

The main purpose of Article 11 is to ensure that controllers don't retain personal data just to support compliance with the GDPR (Art 11(1) GDPR) even if they don't need the data to achieve the purposes of the processing.

If this is the case, Art. 11(2) GDPR waives certain requirements of the GDPR in the case where controllers can demonstrate that they are not in a position to identify data subjects. This can be the case for pseudonymization. To better understand when this is the case, the following gives a short overview. A deeper discussion of the argument can be found in the extended analysis of the topic (see https://uldsh.de/PseudoAnon).

The following analysis is based on the distinction of three kinds of additional information that are required by the purposes of processing:

(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. Referring to Figure 24 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 In this case, the purposes **do not require to *re-identify*** a data subject based on its *pseudonymous handle*. Referring to Figure 24 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*. Referring to Figure 24 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 the following table.

| | | **(i)** | **(ii)** | **(iii)** | |
|---|---|---|---|---|---|
| ***Type of data*** | **identified data** | **strictly pseudonymous data** | **strictly pseudonymous data** | **strictly pseudonymous data** | **anonymous data** |
| **Split-off additional information kept by controller** | N/A | **bi-directional additional information** | **one-directional additional information** | **None** | N/A |
| Is personal data? | yes | Yes | Yes | Yes | no |
| Potential identification of pseudonymous data | direct | indirect<br>(with *additional information* kept by the controller or external) | indirect<br>(with *additional information* external to the controller) | indirect<br>(with *additional information* external to the controller) | not possible<br>(by any actor with means reasonably likely to be used now and in the future) |
| Does the condition of Art. 11(2) apply? | no | No | No | Yes | (yes) |
| Can controller identify data subject autonomously?[120] | yes | yes | No | No | no |
| Can data subject provide suitable additional information to be identified? | N/A | generally yes<br>(typically a unique handle that matches to lookup-based additional information) | generally yes<br>(typically a unique handle as input in the formula-based additional information) | yes, sometimes<br>(unique combination of attributes or *pseudonymous credential*[121]) | no, never |
| Does controller need to implement data | yes | Yes | Yes | yes, unless no single data subject can | no |

[120] "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).

[121] Pseudonymous credentials are described in the extended analysis of the argument. Note that to issue pseudonymous credentials is not required by the GDPR.

| | | | | | |
|---|---|---|---|---|---|
| subject rights | | | | present suitable additional information | |

Based on this analysis, a controller is not in a position to identify a data subject when:

- The controller stores (one- or bi-directional) *split-off additional information* and the data subject can neither provide
  - trusted identity data that matches the input side of the split-off additional information,
  - a *pseudonymous credential*, previously issued by the controller[122], nor
  - a trusted (combination of) value(s) that uniquely matches the *pseudonymized data*.
- The controller stores no split-off additional information and
  - a *pseudonymous credential*, previously issued by the controller, nor
  - a trusted (combination of) value(s) that uniquely matches the *pseudonymized data*.

## 2.6 Anonymization

*Bud P. Bruegger (ULD)*

*Acknowledgements: Parts of the underlying research and writing have been conducted as part of the Forum Privatheit project which has received funding from the German Federal Ministry of Education and Research (BMBF). Discussions with and feedback on an extended version from Marit Hansen, Harald Zwingelberg and Benjamin Bremert, all ULD, and Jessica Schroers, KU Leuven, were very helpful and helped improve the content. Jessica Schroers also reviewed this shortened version of the document.*

*The final version of this section was validated by Hans Graux, guest lecturer on ICT and privacy protection law at the Tilburg Institute for Law, Technology, and Society (TILT) and at the AP Hogeschool Antwerpen. President of the Vlaamse Toezichtscommissie (Flemish Supervisory Committee), which supervises data protection compliance within Flemish public sector bodies.*

[122] Pseudonymous credentials are described in the extended analysis of the argument.

This section describes the GDPR's concept of *anonymization* and how to implement it. It is concerned with rendering identification by any actor and under any realistic circumstances impossible, now and in the future. *Anonymization* is concerned with preventing both direct and indirect identification.

*Anonymization* contrasts with *pseudonymization* which is mostly concerned with direct identification and solely in the controlled environment of the controller's (or joint controllers') processing activity (or activities).

***Anonymization* in a nutshell:**

- Data protection is a fundamental right (see Art. 8 of European Charter of Fundamental Rights).
- The GDPR implements this right by defining the safeguards necessary to protect data subjects at risk through a processing activity.
- *Anonymous data* is data that poses no risk to data subjects.
    - The GDPR therefore does not apply to *anonymous data*.
    - The risk is considered to be absent when *identification* of data subjects is not or no longer possible.
- Data that permits identification is not anonymous
    - even if the risk of identification is small or cannot be perceived, and
    - independently of whether a controller has attempted to anonymize with significant effort and by following the state of the art.
- There is no known test to determine whether data is indeed anonymous.
- Most data sets likely have residual risk that at least partial identification will be possible in the future with newly available additional information, methodology, and computing technology. The term *presumed anonymous data* captures this.
- Some researchers believe that *anonymous data* that is still useful does not exist.
- *Presumed anonymous data* with a residual risk of identification is not (truly) *anonymous* but is *personal data* that is subject to application of the GDPR.
    - Attempted anonymization significantly reduces the risk for data subjects.
    - The GDPR takes a risk-based approach that requires implementing measures and safeguards in proportion to the risk.
    - The most important measure for *presumed anonymous data* with residual risk is *confidentiality*.
    - The best practice of sharing such data is through a contractual agreement with the recipient that passes on certain obligations from the GDPR. (This is very similar to a contractual agreement between controller and processor).

### 2.6.1 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*.

*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:

"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." [123]

[123] The following summary by Hans Graux provides further legal background on the concept:

"These criteria were examined in greater detail in the so-called Breyer decision of the European Court of Justice[123]. In the Breyer case, the applicant requested public authorities to delete a part of their access logs relating to their public websites. He argued that they contained his IP address as a result of his prior use of the websites, and that the IP addresses constituted personal data. The Court affirmed that IP addresses could indeed be qualified as personal data, even if they are dynamic, and even taking into account that identification would require cooperation of the ISPs (who can normally trivially link IP addresses to subscribers).

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, it must also hold in the future. Hence, 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).

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 in the data with means reasonably likely to be used now or in the future.

---

The Court also stressed that, in order to make this assessment in a specific case, "*it must be determined whether the possibility to combine a dynamic IP address with the additional data held by the internet service provider constitutes a means likely reasonably to be used to identify the data subject. [...] [I]n particular, in the event of cyber attacks legal channels exist so that the online media services provider is able to contact the competent authority, so that the latter can take the steps necessary to obtain that information from the internet service provider and to bring criminal proceedings. Thus, it appears that the online media services provider has the means which may likely reasonably be used in order to identify the data subject, with the assistance of other persons, namely the competent authority and the internet service provider, on the basis of the IP addresses stored".* Quoting the Advocate General, the Court also opined that such means would not be available "*if the identification of the data subject was prohibited by law or practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant*".

The Breyer case is occasionally referenced as a hallmark decision that introduced a risk based approach to deciding the legal qualification of personal and non-personal data. Only if the risk of identification was found to be 'insignificant', would data be qualified as purely non-personal. In practical terms, its impact was to significantly increase the awareness of the complexity of the assessment of data: after Breyer, it was no longer sufficient to stress that identification would not normally happen, or that it would require significant efforts, or access to third party data sources. If means existed for the controller that might be likely reasonably to be used for identification, the data should be considered as personal data, and the GDPR would thus need to comply with. As a result, the reach of data protection law was perceived as significantly broader post-Breyer."

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[124] 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.

### 2.6.2 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.

| Definition: ***Anonymous data*** | Definition: ***Strictly pseudonymous 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**. | *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. |

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 as well as the **future beyond**, *strictly pseudonymous data* limits addresses considers only the time of processing. In other words, while *anonymous* uses an open temporal horizon, *strictly pseudonymous* uses a limited temporal horizon.

[124] 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).

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.

Note that *anonymous data* are also *strictly pseudonymous* since the requirements for being *strictly pseudonymous* are a subset of the requirements for being *anonymous*.

The following two figures illustrate the difference between *strictly pseudonymous* and *anonymous*. First, Figure 25 shows the case of *pseudonymization* where the facilitating elements of the environment are shown in green. Then, Figure 26 shows the case of *anonymization* with the more demanding elements highlighted in red.

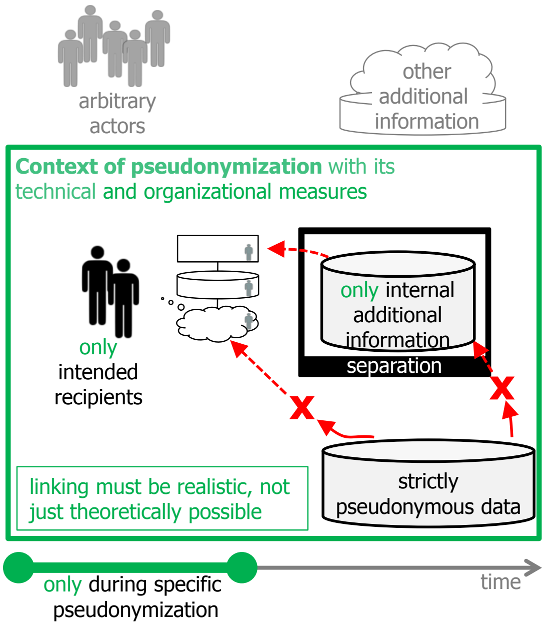

arbitrary actors
other additional information
Context of pseudonymization with its technical and organizational measures
only intended recipients
only internal additional information
separation
strictly pseudonymous data
linking must be realistic, not just theoretically possible
only during specific pseudonymization
time

Figure 25: Data that is pseudonymous in the context of a specific pseudonymization (i.e., processing activity).

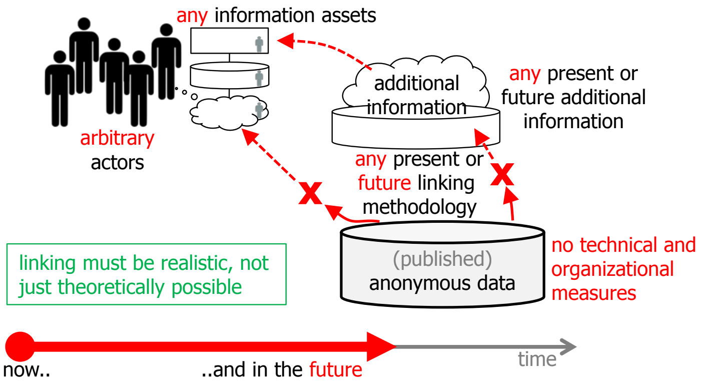

any information assets
arbitrary actors
additional information
any present or future additional information
any present or future linking methodology
linking must be realistic, not just theoretically possible
(published) anonymous data
no technical and organizational measures
now..
..and in the future
time

Figure 26: Anonymous data.

### 2.6.3 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 .

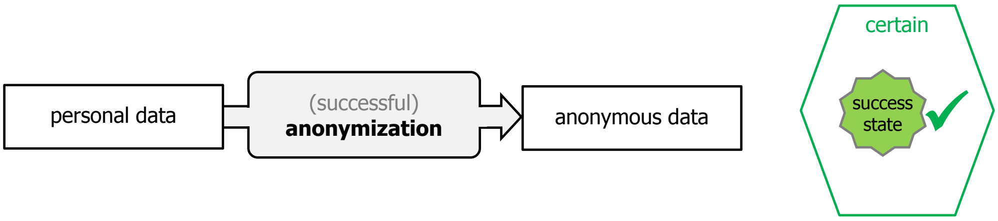


Figure 27: Anonymization.

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

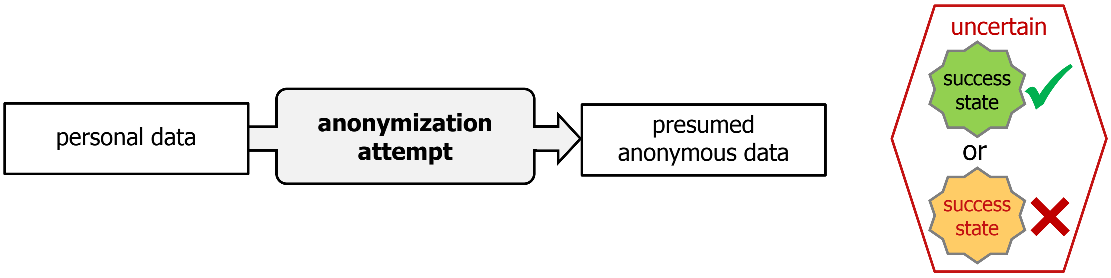


Figure28: Attempted anonymization.

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*. In fact, 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*.

### 2.6.4 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 2.4.3 above). The reduction is considered sufficient, when the “success state” of no longer being able to identify data subjects has been reached.

Since *data pseudonymization* is also implemented by transformations which reduce the identification potential, Figure 29 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”.

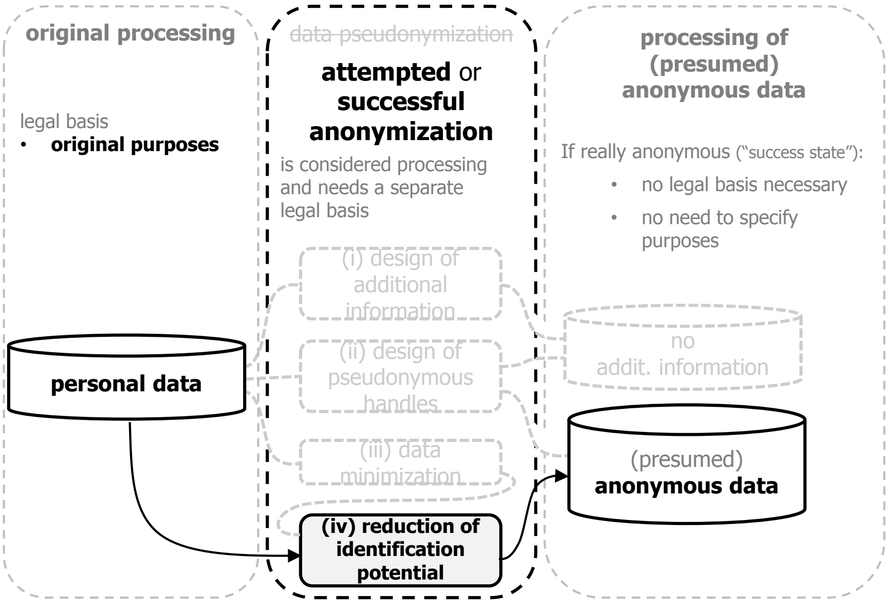

original processing
legal basis
• original purposes
personal data
~~data pseudonymization~~
attempted or successful anonymization
is considered processing and needs a separate legal basis
(i) design of additional information
(ii) design of pseudonymous handles
(iii) data minimization
(iv) reduction of identification potential
processing of (presumed) anonymous data
If really anonymous (“success state”):
• no legal basis necessary
• no need to specify purposes
no addit. information
(presumed) anonymous data

Figure 29: 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 2.4.3.4 above describes how the available transformations reduce the identification potential 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*.

### 2.6.5 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[125], Natasha Lomas[126], Rocher et al.[127] and Dwork et al.[128].

Some kinds of data have been found to be very difficult to anonymize. Most prominently, this holds for location data[129]. Here, even a generalization to country level may not be sufficient[130]. Also, to reduce the identification potential of data, transformation that reduces 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 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" [131]. 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)[132].

### 2.6.6 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. Whether the "success state" applies often depends on the possible external actors, their know-how about re-

[125] 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).

[126] 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).

[127] 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).

[128] 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 (last visited 8/4/2021).

[129] 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).

[130] 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 (last visited 12/8/2021).

[131] 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).

[132] Cynthia Dwork, Introduction: The Definition of Differential Privacy, Institute for Advanced Study, Four Facets of Differential Privacy, November 12, 2016, https://youtu.be/lg-VhHlztqo?t=180 (last visited 8/4/2021).

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.

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 at 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

n 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, one of the following cases could occur:

- *Identification is already possible,*
- *identification will eventually be possible,* or
- the data is "truly" *anonymous.*

In the former two cases, the data are *personal* and the GDPR is applicable[133]; 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 30 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 data as *anonymous* or *personal data*, respectively.

| **Data:** | **Treat as anonymous data** | **Treat as personal data** |
|---|---|---|

[133] This is because the concept of "the means reasonably likely to be used" is inherently a forward looking criterion.

| | | |
|---|---|---|
| **“truly”** ***anonymous*** | [correct classification]<br>**GDPR-compliant**<br>**Obligations according to GDPR**:<br>in some cases, a Data Protection Impact Assessment (DPIA) is required before anonymization[134] | [incorrect classification]<br>**GDPR-compliant**<br>(extra effort is allowed)<br>**Obligations according to GDPR**:<br>none, but implementation of measures insures against consequences of classification error. |
| ***identification will eventually be possible*** | [incorrect classification]<br>**GDPR violation**<br>**Potentially irreparable damage for data subjects**<br>**Obligations according to GDPR**:<br>Mandatory damage control, possible termination of processing, consequences of GDPR violation and potential liability claims | [correct classification]<br>**GDPR-compliant**<br>**Obligations according to GDPR**:<br>Implementation of technical and organizational measures. |
| ***identification is already possible*** | [incorrect classification]<br>**GDPR violation**<br>**Potentially irreparable damage for data subjects**<br>**Obligations according to GDPR**:<br>Mandatory damage control, possible termination of processing, consequences of GDPR violation and potential liability claims | [correct classification]<br>**GDPR-compliant**<br>**Obligations according to GDPR**:<br>Implementation of technical and organizational measures. |

**Figure 30: The different options available to controllers to deal with *presumed anonymous* data.**

[134] 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.lda.bayern.de/media/dsfa_muss_liste_dsk_de.pdf (last visited 12/8/2021).

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) What is the mandatory damage control?

(4) How substantial is the effort of treating presumed anonymous data as being personal when there is any doubt?

#### 2.6.6.1 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 processes 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 in favor 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. As a result, these data subjects may suffer highly adverse consequences at their workplace, in their career, as well as their relationships.

It is also evident here that once such damage is done, it is irreversible and beyond remediation.

#### 2.6.6.2 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.

The extended version of this analysis (see https://uldsh.de/PseudoAnon) provides reasons why this situation could be considered to be a data breach according to Art. 4(12) GDPR. It is a cautious course of action for controllers to treat it as such.

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.

#### 2.6.6.3 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. The following provides a short summary of the extended analysis of this topic (see https://uldsh.de/PseudoAnon).

Since the processing needs to comply with the GDPR, all its requirements must be met as rapidly as possible or else any further processing has to be terminated.

The following summarizes the kinds of actions that are required to contain the damage. It looks at past and present processing operations:

*Past processing operations*:

- Create retrospective compliance where possible (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).
- 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[135]).

### 2.6.6.4 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 briefly looks at what exactly has to be done to "play it safe" and treat *presumed anonymous* data as *personal data*.

The effort is usually quite contained, at least for organizations who are already familiar with the requirements of data protection[136].

The most significant difference as compared to treating the data as *anonymous* is that confidentiality is required. Publication of the data, i.e., disclosure to arbitrary third-party recipients, is evidently the contrary of confidentiality. In fact, the disclosure to selected recipient is possible when there is a valid legal basis for such disclosure.

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 practice to propagate the necessary obligations and limitations to recipients is through 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 U.S. example[137] 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)[138]. 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[139].

Such a contractual agreement between a controller and a third-party recipient could regulate the following:

[135] An example for such decision-making would be the refusal of a credit or service, or the denial of a right.

[136] For example, such organizations already have knowledge of their obligations and have appointed a DPO (if required).

[137] https://www.hcup-us.ahrq.gov/team/NationwideDUA.jsp (last visited 10/5/2021).

[138] https://www.hcup-us.ahrq.gov/ (last visited 10/5/2021).

[139] See https://aircloak.com/the-five-private-eyes-part-1-the-surprising-strength-of-de-identified-data/ under HCUP, (last visited 10/5/2021).

- 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 thereof 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.

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.

## 2.7 Data Protection and Scientific Research

*Pilar Nicolás Jiménez*[140] *(UPV/EHU), Mikel Recuero Linares (UPV/EHU)*

*This part of the Guidelines was reviewed by Rossana Ducato*

*This part of The Guidelines has been reviewed and validated by Marko Sijan, Senior Advisor Specialist, (HR DPA)*

### 2.7.1 Key Points

[140] This section incorporates some references extracted from a book chapter by the author, originally published in Spanish: *Comentarios al Reglamento General de Protección de Datos y a Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales* (Antonio Troncoso Reigada, Dir.), Thomson Reuters Aranzadi, 2020.

- The GDPR is conscious of the paramount relevance that the processing operations with purposes of archiving in the public interest, scientific or historical research purposes or statistical purposes may imply.
- Therefore, the Regulation envisages a special and favorable regime in an attempt to ensure that the data protection rules do not constitute a major hurdle to the processing operations for the referred purposes.
- In this regard, its necessity for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes is expressly laid down as a condition for the processing of special categories of personal data.
- The text also establishes a flexible regime for long-term data storage and a presumption of compatibility for secondary or further purposes.
- In addition, limitations, exceptions, or derogations are provided, *inter alia*, to the rights to information, access, rectification, restriction of processing, object and, concerning the archiving purposes in the public interest, to the right of notification and portability.
- In order to strike the right balance with data subject's rights and interests, the Regulation requires the adoption of appropriate safeguards in accordance with Article 89 and, in certain situations, also further development by Union or Member State law.

### 2.7.2 **Introduction**

As the European Data Protection Supervisor (EDPS) highlighted, "the European Commission has defined the objectives of the EU's research and innovation policies to be 'opening up the innovation process to people with experience in fields other than academia and science', 'spreading knowledge as soon as it is available using digital and collaborative technology' and 'promoting international cooperation in the research community'".[141] These purposes are not in conflict with data protection. Indeed, data protection rules should not be an obstacle to freedom of science pursuant to Article 13 of the Charter of Fundamental Rights of the EU (CFREU). Rather, these rights and freedoms must be carefully assessed and balanced, resulting in an outcome which respects the essence of both.[142]

Indeed, the intention behind our current data protection legislation is to harmonize data processing with scientific research purposes.[143] This intention is clearly linked to Article 179(1) of the Treaty on the Functioning of the European Union (TFEU) for achieving a European Research Area. In line with this, the General Data Protection Regulation (GDPR) has introduced a new framework aimed at enabling data processing for archiving purposes in the public interest, historical and scientific research purposes

[141] EDPS, A Preliminary Opinion on data protection and scientific research, 2020, p. 10. At: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202003_healthdatascientificresearchcovid19_en.pdf Accessed: 15 January 2020.

[142] EDPB, Guidelines 03/2020 on the processing of data concerning health for the purpose of scientific research in the context of the COVID-19 outbreak. Adopted on 21 April 2020, p. 5. At: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202003_healthdatascientificresearchcovid19_en.pdf Accessed 23 April 2020.

[143] Recital 159 GDPR.

or statistical purposes that goes beyond that provided by Directive 95/46/EC.[144] The core of this new regulation is Article 89 of the GDPR, which is accompanied by many other references throughout the whole text that complete it. These can be found both in the part of the GDPR that includes the decisive criteria for its interpretation (recitals), and in some specific provisions[145]. On the basis of those recitals, some preliminary ideas should be highlighted.

First, Recital 157 states that by coupling information from registries, including different types of data corresponding to a lot of individuals, researchers can obtain "new knowledge of great value with regard to widespread medical conditions such as cardiovascular disease, cancer and depression". As a consequence, "research results can be enhanced, as they draw on a larger population". These tools can contribute to improving research policies and, consequently, the population's quality of life. These benefits mean that the processing of data for these purposes by researchers is reasonable, provided that the rights of the subjects are guaranteed. This establishes a conception of research as a process that pursues a social benefit, in the short, medium or long term, considered in a very broad way (improvement of the quality of life) but, at the same time, limiting that activity to this specific purpose. Furthermore, recital 159 specifies that "to meet the specificities of processing personal data for scientific research purposes, specific conditions should apply in particular as regards the publication or otherwise disclosure of personal data in the context of scientific research purposes".

The second issue to be addressed is the specific nature of consent as a requirement for its validity, which has some particularities when the purpose of the processing is scientific research. Indeed, Article 4 of the GDPR states that consent "means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her". However, recital 33 states that "it is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection".

However, it is common that during a project, approaches not initially foreseen may emerge, or that, upon completion of the project, the conclusions open doors to other related projects. Furthermore, researchers and teams are often specialized in an area or line of research developed from specific projects, and the data may remain useful or necessary for long periods of time[146]. As a response, institutional models – such as biobanks – have emerged, functioning as intermediaries between subjects and researchers. The purpose of collecting these data is to store them for when they may be

[144] 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).

[145] See, *inter alia*: Article 5(1)(b) for compatible purposes; Article 5(1)(e) concerning storage limitation; Article 9(2)(j) as a derogation for the processing of special categories of data; Article 14(5)(b) concerning transparency and information; Article 17(3)(d) referring to the right to erasure; or Article 21(6) for the right to object.

[146] In this regard, see also Article 5(1)(e) of the GDPR that allows personal data to be stored for longer periods insofar they are processed solely "for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1).

required, without knowing, in principle, which research project, or projects, will process them. In view of this reality, Recital 33 states that "data subjects should be allowed to give their consent to certain areas of scientific research" even though "data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose". Different options and consent are therefore allowed to varying extents provided that they are, as the recital recalls, "in keeping with recognized ethical standards for scientific research".

A third point that deserves attention is that contained in Recital 50, which refers to the so-called compatibility of purposes[147], i.e., "processing of personal data for purposes other than those for which the personal data were initially collected". This is a term used in cases where the personal data, intended to be used for research purposes, were initially collected or processed for a different purpose, but it can be legitimately processed for further new (compatible) purposes. In addition, further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes are *ex lege* considered compatible lawful processing operations. This means that no consent of the data subject nor other legal basis are required for this further purpose, under the conditions to be described later. This option is of utmost importance for scientific research because it can facilitate access to a huge amount of data without the need to re-contact the data subjects.

Finally, it is necessary to mention Recital 53, which takes up the purpose of the GDPR concerning the establishment of harmonized conditions for the processing of special categories of personal data for health-related purposes (in particular, in the context of the management of health or social care services and systems). Furthermore, it states that "Union or Member State law should provide for specific and suitable measures so as to protect the fundamental rights and the personal data of natural persons", while declaring that "Member States should be allowed to maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health." However, introduced measures "should not hamper the free flow of personal data within the Union when those conditions apply to cross-border processing of such data".

### 2.7.3 Notions in the context of the EU regulatory framework

#### A. Notion of "purposes of archiving in the public interest"

Archives in the public interest are understood to be those of public or private bodies that hold records of public interest and which, pursuant to Union or Member State law, have a legal obligation to acquire, preserve, appraise, arrange, describe, communicate, promote, disseminate and provide access to records of enduring value for general public interest[148]. Nevertheless, it does not apply to deceased persons' data (see "Personal Data", in Part II of these Guidelines, section "Main Concepts").

#### B. Notion of "scientific research"

[147] In this regard, see also Article 5(1)(b) of the GDPR.

[148] Recital 158 of the GDPR.

Scientific research is an overly broad term that generally refers to the search for knowledge, through a certain methodology, in any area of human knowledge. The GDPR does not include a definition of "scientific research" as such but introduces a series of considerations that allow us to define its main characteristics. Firstly, "scientific" research is different from "historical research purposes" and "statistical purposes". Furthermore, it covers different fields e.g., research in the life sciences related to human health, but also the social sciences (recitals 157 and 159). It must bring "benefits", at least potentially. This expectation justifies a unique regime that allows exceptions and derogations of certain rights (Art. 89.2)[149]

Within this framework, the GDPR undertakes a broad interpretation of scientific activity, including "technological development and demonstration, fundamental research, applied research and privately funded research" (recital 159). This broad conception includes research projects with publishable results and other analytical studies, without excluding privately funded research or research funded by profit-seeking commercial companies. However, it also contains certain limits, some criteria that make it possible to determine the extent to which the exceptions provided for throughout the GDPR can be applied in a scenario of increasing data analysis procedures. However, the Regulation remains ambiguous about what parameters an activity or processing operation must meet in order to be considered "scientific research". The EDPS has, in an attempt to shed some light on this, alluded to the following parameters in its Preliminary Opinion on data protection and scientific research[150]:

- The activity must contribute to the increase of knowledge (scientific research in the strict sense) or the use of knowledge for the production of devices, materials, services, processes or products (technological development and demonstration).
- The activity must be developed under certain quality standards (professional, methodological and institutional), "including the notion of informed consent, accountability and oversight"[151].
- "The research is carried out with the aim of growing society's collective knowledge and wellbeing, as opposed to serving primarily one or several private interests".[152]

According to this perspective, scientific research, for the purposes of the GDPR, covers the activity of both generating and applying knowledge and excludes activity that does not present a guarantee of rigour in its development. Thus, scientific research requires research projects to be "set up in accordance with relevant sector-related methodological and ethical standards, in conformity with good practice".[153] The

[149] Recital 157 GDPR.

[150] EDPS, A Preliminary Opinion on data protection and scientific research, 2020, p. 12. At: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202003_healthdatascientificresearchcovid19_en.pdf Accessed: 15 January 2020.

[151] Ibid.

[152] Ibid.

[153] EDPB, Guidelines 05/2020 on consent under Regulation 2016/679, adopted on 4 May 2020, v1.1., p. 30. Available at:

procedures that allow the adequate evaluation of these parameters, which may vary from case to case, will represent for the processing of the data in the sense of Article 89.1.

It is important to underline that teaching[154] cannot be considered a scientific activity, even if it is aimed at training professionals in this sector. Consequently, given that the GDPR does not include any mention of it, the processing of data for this purpose is subject to the general regime, which can lead to many dysfunctions in practice.[155]

### C. Notion of "historical research"

The GDPR applies this description to data processed for the purposes of historical research. This is a broad notion that includes both historical research itself and research for genealogical purposes[156]. However, it does not apply to research carried out with deceased persons' data.

### D. Notion of "processing for statistical purposes"

Statistical purposes mean any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results[157]. However, the resulting data must be non-personal data (aggregate data), and it is further required that neither this result nor the personal data are used in support of measures or decisions regarding any particular natural person.

In addition, once again, Union or Member State law, within the limits of the GDPR, should determine most of the practical and particular aspects of the processing (what data is considered as statistical content, control of access, and appropriate measures to safeguard the rights and freedoms of the data subject and for ensuring statistical confidentiality, etc.).

## 2.7.4 Which data does Article 89 cover?

One other relevant point of discussion is the nature of data for which processing requires appropriate safeguards and may justify exceptions and derogations to the rights of the subjects.

There is no doubt that Article 89 comprises all categories of personal data, which also includes the processing of special categories of personal data, provided that the conditions for the processing of these latter are met.

Thus, the secondary use of data for archiving purposes, for scientific research or historical research, or for statistical purposes (Art. 5), must be supported by the guarantees referred to in Art. 89 when, for example, analysis or cross-checking with other data highlights information of a sensitive nature. Therefore, when applying the

https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf Accesed 16 September 2021.

[154] "Teaching" must not be identified with "academic expression" in the context of Art. 85 GDPR.

[155] See, about "academic expression", EDPS, p. 10.

[156] Recital 160 GDPR.

[157] Recital 162 GDPR.

regime of Art. 89, the context of the processing, its implications and the nature of the data are of paramount importance.

### 2.7.5 Purpose compatibility

According to Article 5(1)(b), the further processing for the purposes of archiving in the public interest, scientific or historical research purposes or statistical purposes is compatible even if the data were collected initially for other purposes (provided that technical and organizational measures are in place that ensure respect for the rights and freedoms of the data subject). However, it remains under discussion whether other provisions may apply e.g., the compatibility test under Article 6(4) of the GDPR.

However, in relation to special categories of data, Article 9(2) (j) explicitly mentions that processing must be "based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject".

This apparent legal issue requires an interpretative effort that could resolve the matter in two ways. First, since Article 5 does not refer to special categories of personal data, it could be understood as limited to cases where no such information is used. If we were to speak of personal data of these categories, Article 9, which is more specific, would apply.

The second solution is based on an interpretation of Article 5 merely as general principles, and in light of recital 50, which outlines a number of conditions for secondary use, representing the requirement of enhanced self-monitoring by the controller, as well as a "reasonable expectation" on the part of the data subject that this secondary processing can take place. In addition, Art. 6(4) establishes a number of criteria to determine the compatibility of a processing operation with the (different) purpose for which the personal data were collected, which should also be taken into account in these cases: "a) any link between the purposes for which the personal data have been collected and the purposes of the intended further processing; b) the context in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller; c) the nature of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10; d) the possible consequences of the intended further processing for data subjects; e) the existence of appropriate safeguards, which may include encryption or pseudonymization" (see "Identification", "Pseudonymization" and "Anonymization" in Part II of these Guidelines, section "Main Concepts"). Therefore, it seems that Articles 5, 6 and 9 should be read and interpreted together[158].

[158] EDPS, A Preliminary Opinion on data protection and scientific research, 2020, p. 23. At: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202003_healthdatascientificresearchcovid19_en.pdf Accessed: 15 January 2020.

### 2.7.6 Conceptual issues: legal basis for processing.

As far as the legal basis for processing is concerned, it is relevant to distinguish between categories of data:

- Processing of personal data ('non-sensitive'). The legal bases for the processing are those set out in Article 6 of the GDPR (see "Lawfulness, Fairness and Transparency" in Part II section "Principles" of these Guidelines). This means that every processing of personal data must necessarily rely on any of the legal basis pursuant Article 6(1):
  a) Consent of the data subject (art. 6.1 a).
  b) Contract (art. 6.1 b).
  c) Legal obligation (art. 6.1 c).
  d) Vital interests (art. 6.1 d).
  e) Public task or public interest (art. 6.1 e).
  f) Legitimate interests (art. 6.1 f).
- Processing of special categories of personal data ('sensitive personal data'). The processing of those categories of data included in Article 9 is forbidden unless a specific legitimate basis from those in Article 9(2) is identified.[159] Article 9 requires further legitimation, added to those in Article 6. Between these legal bases, processing is not banned, if, among other things:
  a) "the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject". Article 9(2) letter a).
  b) it is "necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject".[160]

  In addition, Article 9 (4) reads: "Member States may maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health." This possibility does not, however, imply that the content of paragraph (2)(j) of Article 9 should be rendered ineffective. Again, researchers should always ask their DPOs for advice about the applicable national regulatory framework.

[159] EDPB, Opinion 3/2019 concerning the Questions and Answers on the interplay between the Clinical Trials Regulation (CTR) and the General Data Protection regulation (GDPR) (Art. 70.1.b)) Adopted on 23 January 2019, pp. 8–9. At: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_opinionctrq_a_final_en.pdf Accessed: 20 May 2020.
This document describes several possibilities that combine Articles 6 and 9: The lawful grounds for processing can be derived from legal obligations of the controller and which fall within the legal basis of Article 6(1)(c) in conjunction with Article 9(1)(i); or the public interest under Article 6(1)(e) in conjunction with Article 9(2), (i) or (j); or the legitimate interests of the controller under Article 6(1)(f) in conjunction with Article 9(2)(j); or under specific circumstances, when all conditions are met, data subject's explicit consent under Article 6(1)(a) and 9(2)(a).

[160] Article 9(2)(j).

### 2.7.7 Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes and the right to information

Where personal data have not been obtained from the data subject and the processing is carried out for purposes of archiving in the public interest, scientific or historical research purposes or statistical purposes, derogations to the right to information are foreseen by the Regulation. In close relation to the two previous points, and to facilitate the availability of data for those purposes, Article 14(5)(b) of the GDPR provides that the provisions of paragraphs 1 to 4 (describing the information that the data controller must transfer to the data subject) shall not apply when "the provision of such information proves impossible or would involve a disproportionate effort, in particular for processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, subject to the conditions and safeguards referred to in Article 89(1) or in so far as the obligation referred to in paragraph 1 of this Article is likely to render impossible or seriously impair the achievement of the objectives of that processing. In such cases the controller shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available."

Therefore, as it can be inferred from the literacy of the provision, further development by Union or Member State's law is not required here in order to apply this derogation.

### 2.7.8 Derogations to certain rights of the data subjects pursuant Article 89

Article 89(2) of the GDPR states: "Where personal data are processed for scientific or historical research purposes or statistical purposes, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18 and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes." Together with Article 14(5)(b), this clause introduces several derogations regarding data subject rights (see "Data Subject Rights" section within Part II of these Guidelines), namely:

- *Right of access (Article 15 GDPR):* According to Article 89, it is possible to limit the right of access to data subjects. This limitation covers both personal data that was processed for the research and personal data that was obtained as a result of the analysis or procedures developed. In the biomedical field, for example, this concerns any results obtained from body examinations or procedures, analysis of their samples or data, etc.

- *Right to rectification (Article 16 GDPR):* The right to have inaccurate data rectified or completed is not of great significance in scientific research (it may be more relevant, for example, in historical research). Neither is its limitation. The methodology of scientific research requires accuracy and reliability of the information being handled in order that solid conclusions are obtained, so it will be in its own interest to demand such accuracy;

- *Restriction of processing (Art. 18 GDPR):* Restriction of processing "means the marking of stored personal data with the aim of limiting their processing in the future" (Article 4(3) GDPR). Personal data whose processing is limited are not deleted and are kept for different purposes, but cannot be used or transferred beyond that scope. In the framework of an investigation, the exercise of this right could hinder the continuity of the investigation or the publication of results in its first phase (limitation of the continuity of its use). This is why this derogation makes sense.

- *Right to object (Art. 21 GDPR):* The right of objection allows the data subject whose personal data is being processed pursuant to any legal grounds other than consent to object to the processing. This possibility is the basis of so-called opt-out systems (in which consent to the use of data for research purposes is presumed), and fundamental for cases in which consent to processing is not required (Articles 5 and 9 GDPR). Raising exceptions to this right has important consequences for the autonomy of the data subjects, since it may imply that the data are used against their will. Justifying these exceptions as obstacle they may represent for research, would be fairly simple in any case where such data are relevant for research.

**Example: Research on rare diseases**

Research on rare diseases often relies on personal data obtained from a quite small number of data subjects (due to the pure nature of rare diseases). Therefore, if a significant number of individuals participating in the research decide to exercise their rights to restriction and/or objection, the representativeness and reliability of the research data might be significantly undermined as a consequence. Furthermore, researchers might face serious issues in terms of publishing, since they could not provide those data to the publisher. Therefore, under such circumstances, the controller could use the derogations to those rights settled by Article 89.

Controllers shall always keep in mind that "any derogation from these essential data subject rights must be subject to a particularly high level of scrutiny in line with the standards required by Article 52(1) of the Charter". As a result, derogations under GDPR Article 89(2) are only possible if the conditions and safeguards required under Article 89(1) are satisfied.

Furthermore, under Article 89(2), derogations can be applied only "in so far as" the rights to be derogated from are "likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes".[161] Lastly, controllers must consider that "the fact that putting in place technical and organizational measures to provide access and other rights

[161] EDPS, A Preliminary Opinion on data protection and scientific research, 2020, p. 21. At: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202003_healthdatascientificresearchcovid19_en.pdf Accessed: 15 January 2020.

to individuals may require financial and human resources is by itself not a valid justification to derogate from the rights of individuals under the GDPR".[162]

Finally, as regards only to the data processed for archiving purposes in the public interest, Union or Member State law may provide, in addition to those above-mentioned, for derogations to the right of notification regarding rectification, erasure or restriction of processing (Article 19) and to the right of portability (Article 20)[163]. Once again, this requires that the exercise of these rights may render impossible or seriously impair the achievement of the specific purposes and that such derogations may be, as a result, necessary for the fulfilment of those purposes.

*Derogations to the right to erasure or right to be forgotten*

According to Article 17(3)(d), this right shall not apply to the extent that processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) in so far as it is likely to render impossible or seriously impair the achievement of the objectives of that processing.

Similarly, derogations to the right to erasure will directly apply, without the need of further development by Member States.

### 2.7.9 Storage limitation

According to Article 5(1)(e) of the GDPR, personal data should be "kept in a form which permits identification of data subjects for no longer than is necessary" (see "Storage limitation principle" within Part II section "Principles" of these Guidelines). However, the GDPR permits storage for longer periods if the sole purpose is scientific research (or archiving in the public interest, historical research or statistical purposes), provided that controllers are allowed to proceed to such processing under an adequate legal basis (storage involves data processing). "The intention of the lawmaker appears to have been to dissuade unlimited storage even in this special regime, and guards against scientific research as a pretext for longer storage for other, private, purposes. If in doubt, the controller should consider whether a new legal basis is appropriate."[164]

Therefore, storage periods should be proportionate to the aims of the processing. "In order to define storage periods (timelines), criteria such as the length and the purpose of

[162] EDPS Opinion on safeguards and derogations under Article 89 GDPR in the context of a proposal for a Regulation on integrated farm statistics, 2017. p.3. At: https://edps.europa.eu/sites/edp/files/publication/17-11-20_opinion_farm_statistics_en.pdf. Accessed: 17 January 2020.

[163] See Article 89(3) of the Regulation.

[164] EDPB, Guidelines 03/2020 on the processing of data concerning health for the purpose of scientific research in the context of the COVID-19 outbreak. Adopted on 21 April 2020, p. 10. At https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf Accessed: 23 April 2020.

the research should be taken into account. It has to be noted that national provisions may stipulate rules concerning the storage period as well."[165]

### 2.7.10 Appropriate safeguards to be adopted pursuant Article 89(1)

Article 89(1) requires that "appropriate safeguards" be applied to the processing of personal data for scientific or historical research or statistical purposes, no matter what the legal basis for processing might be. The purpose of these safeguards is to ensure respect for the principle of minimization of personal data (see "Minimization principle" subsection in the "Principles" section within Part II of these Guidelines. Thus, the first parameter to be analyzed is whether the very conditions for the processing of personal data are met, i.e., the processing of personal data must be necessary to carry out that particular research. Article 89(1) provides that appropriate safeguards "must be reflected in technical and organizational measures", such as pseudonymization. Pseudonymization must be accompanied by other provisions, depending on the risks involved in each project. Controllers should always ensure the implementation of adequate technical and organizational measures aimed at ensuring the protection of data subjects' rights and freedoms. The following are some possible examples of such measures or safeguards:

- Control of access to databases in a manner that such access is only allowed to authorized persons, for approved research, with justified scientific interest, and implemented software solution that allows auditable control access log files.

- Signing of a legally binding commitment between the parties, which includes the conditions of the processing: commitment to confidentiality and non-identification of the data subjects, and use of the data for the specific authorized purpose.

- Implementing security measures for ensuring protection of transfer and storing of data at recipient.

- Ensure transparency of the information provided to the participants.

- Continuous monitoring of the processing conditions over time, which could take the form of transparency measures (publication and accessibility of data management policies) and long-term forecasts (identification of the obligations of the data controller). In relation to this last point, the need to establish clear commitments to monitor the management/handling of personal data by the institution which conducts the research and which could be more specifically entrusted to the corresponding Research Ethics Committee (REC) must be stressed.

- Establishment of a control system external to the investigator that could fall within the competence of the corresponding REC or the management of the research center, which should be involved in the above-mentioned agreement.

Furthermore, researchers should bear in mind that there are other mechanisms provided for in the general GDPR regime that also introduce appropriate measures to the processing of data for research purposes in the sense of Article 89(1), such as the DPIAs or the intervention of the DPOs. Finally, it is interesting to mention that there are

[165] Ibidem, p. 10.

initiatives to promote international codes of conduct and certification mechanisms that may harmonize these safeguards.

### 2.7.11 Further Reading

- EDPS, Opinion on safeguards and derogations under Article 89 GDPR in the context of a proposal for a Regulation on integrated farm statistics, 2017. At: https://edps.europa.eu/sites/edp/files/publication/17-11-20_opinion_farm_statistics_en.pdf
- EDPS, A Preliminary Opinion on data protection and scientific research, 2020. At: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202003_healthdatascientificresearchcovid19_en.pdf.
- EDPB, Opinion 3/2019 concerning the Questions and Answers on the interplay between the Clinical Trials Regulation (CTR) and the General Data Protection regulation (GDPR) (Art. 70.1.b). Adopted on 23 January 2019. At: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_opinionctrq_a_final_en.pdf
- EDPB, Guidelines 03/2020 on the processing of data concerning health for the purpose of scientific research in the context of the COVID-19 outbreak. Adopted on 21 April 2020. At: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202003_healthdatascientificresearchcovid19_en.pdf