AnoMed
Kompetenzcluster Anonymisierung für medizinische Anwendungen

Deliverable 4.9.4

Terminologies on Pseudonymization and Anonymization

Funded by Bundesministerium für Bildung und Forschung

Funded by the European Union NextGenerationEU

https://anomed.de

UAP	4.9.4
Date	19.09.2024
Version	1.0
Status	Final
Distribution (only final version)	PU
Lead Contributors (© by affiliation)	Bud P. Bruegger (ULD)
Additional Contributors (© by affiliation)	
Reviewers	Harald Zwingelberg (ULD)
License	CC-BY 4.0 (only if status is final and distribution PU)

Executive Summary

Two short terminologies on pseudonymization and anonymization were developed in AnoMed. They pursue the larger objective of bridging the gap between science and policy. They attempt to enable AnoMed to contribute to the European strategy of large-scale reuse of data that promises important benefits for our society.

In the case of reuse of personal data, these important benefits are contrasted by the risk of unintended disclosure of personal data. This materialized when supposedly anonymous data can unexpectedly be re-identified. Consequently, to harvest the benefits at minimal risk, policy decisions have to be well-informed. This deliverable describes this process by providing some background on technology transfer.

To contribute to informed policy decisions, AnoMed has analyzed the conceptualization that is implied by both, legal/policy and technical/scientific, texts. This analysis resulted in the identification of certain mismatches between the description of technical artefacts in legal/policy texts and technical reality.

Two terminologies have been developed that policy makers can use and that convey a more harmonized conceptualization that is compatible with the technical understanding. Using these more realistic concepts for describing the policy decisions, it becomes more likely that the outcome can find a successful technical implementation.

The deliverable presents the terminologies and how they incorporate measures to harmonize the conceptualization between the legal/policy and technical/scientific worlds. It also reports on initial dissemination efforts that are crucial to create an impact.

Table of Contents

1 Objective

Terminology is at the core of language. It names the concepts that are used to think about a certain aspect of the world. Different disciplines typically use different languages and thus distinct terminologies. In AnoMed, this holds between technical/scientific disciplines and partners and the policy/legal domain and partners. To foster mutual understanding and dialog, the language barrier between the disciplines has to be removed. A step in this direction is a common terminology that is object of this deliverable.

Figure 1 illustrates the situation. In particular, it shows the position as intermediary of ULD. As a data protection supervisory authority, ULD is clearly positioned in the policy/legal space. With its research department, it also links to the technical/scientific space and is therefore uniquely positioned in the AnoMed project to build bridges between the two disciplines; here in the form of terminologies.

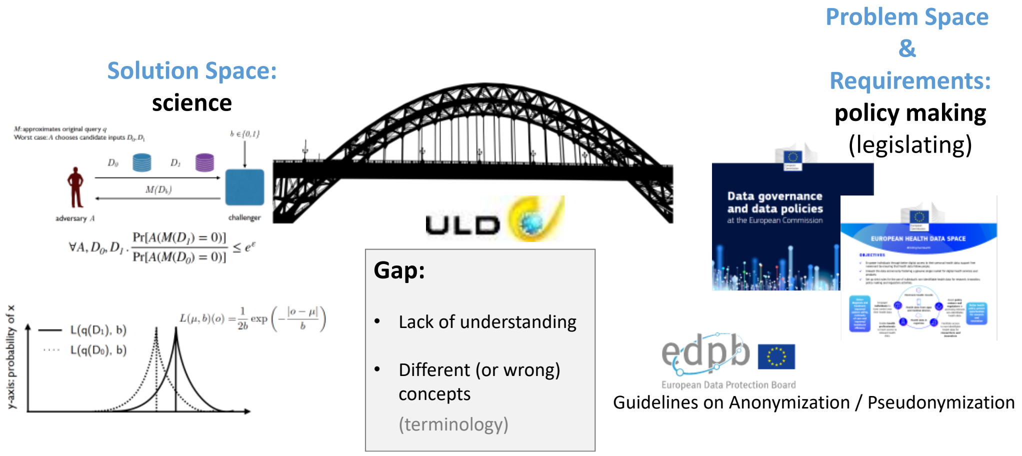

Solution Space:
science
M: approximates original query q
Worst case: A chooses candidate inputs D_0, D_1
$b \in \{0,1\}$
D_0
D_1
$M(D_b)$
adversary A
challenger
$\forall A, D_0, D_1 \cdot \frac{\Pr[A(M(D_1) = 0)]}{\Pr[A(M(D_0) = 0)]} \leq e^{\varepsilon}$
y-axis: probability of x
L(q(D₁), b)
L(q(D₀), b)
$L(\mu, b)(o) = \frac{1}{2b} \exp\left(-\frac{|o-\mu|}{b}\right)$
ULD
Gap:
Lack of understanding
Different (or wrong) concepts
(terminology)
Problem Space
&
Requirements:
policy making
(legislating)
Data governance and data policies
at the European Commission
EUROPEAN HEALTH DATA SPACE
edpb
European Data Protection Board
Guidelines on Anonymization / Pseudonymization

Figure 1: Bridging the gap between science and policy.

A wider picture is illustrated in Figure 2. In particular, it shows that important policy decisions are being made in the context of the European Data Strategy. These decisions happen in an area of rapid advances. They decide on the use of technologies in our society that have a great potential but also significant risks. It is therefore important that the decisions are well informed. For technologies, this means that decision makers possess a realistic understanding of what technologies can achieve and where their limitations are.

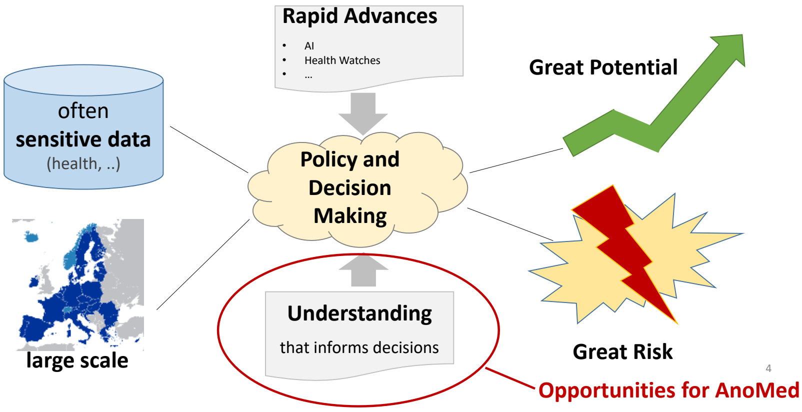


Figure 2: Informed policy decisions require the input of technology and science.

2 Outline

The deliverable is structured as follows. Section 3 provides some background in technology transfer in order to understand the role and purpose of the developed terminologies. Section 4 analyzes the concepts used to talk about anonymization and pseudonymization in the technical/scientific and policy/legal domain. On this basis, Section 5 identifies where conceptualizations are different and where harmonization of the understanding is needed. Section 6 briefly lists some requirements for the terminologies that guided their development. Section 7 presents the two terminologies that were developed for AnoMed in UAP 4.9.4. Section 8 gives a brief overview of initial efforts to disseminate the terminologies outside of the project. Finally, the deliverable concludes with Section 9. Technology Transfer as Background

3 Technology Transfer as Background

This section uses technology transfer as background to better understand the role and purpose of the terminologies developed in AnoMed. Technology transfer is considered for the concrete case of the *European Data Strategy*[1]. It is illustrated in Figure 3.

The left box of the figure shows how the new technologies are introduced in society. The strategy requires technologies such as data analysis and heavily relies on the concepts of anonymization and pseudonymization.

The first step of introducing the technology is a **political strategy**. In this concrete case, the strategy is formulated in a communication from the European Commission (EC) to the Parliament.

Once the strategy has been adopted, it is further implemented in a **series of legal acts** (see also Deliverable D4.9.2). Two that are particular relevant to the present discussion are the *Data Governance Act*[2] (in force) and the *European Health Data Space*[3] (that exists as a proposal by the EC). These acts make major decisions about what technical artefacts need to be implemented and what their properties are. They also restrict the space of architectural options by describing interactions and data flows between different parties. The legislative process of creating and adopting legal acts can be complex and time consuming. After adoption, making changes to a legal act is usually very difficult and therefore mostly avoided.

Once the legal acts are adopted, their **technical implementation** can start. These involve technical decisions at a more detailed level. The process relies on the possibility that all technical artefacts that are described in legal acts can be successfully technically implemented and provide all the properties that are stated in the legal text.

[1] COM/2020/66 final, CELEX 52020DC0066, COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A European strategy for data, 19/2/2020, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066

[2] Regulation (EU) 2018/1724 (Data Governance Act), CELEX 32022R0868, Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32022R0868

[3] COM/2022/197 final, CELEX 52022PC0197, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space, 3/5/2022, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0197

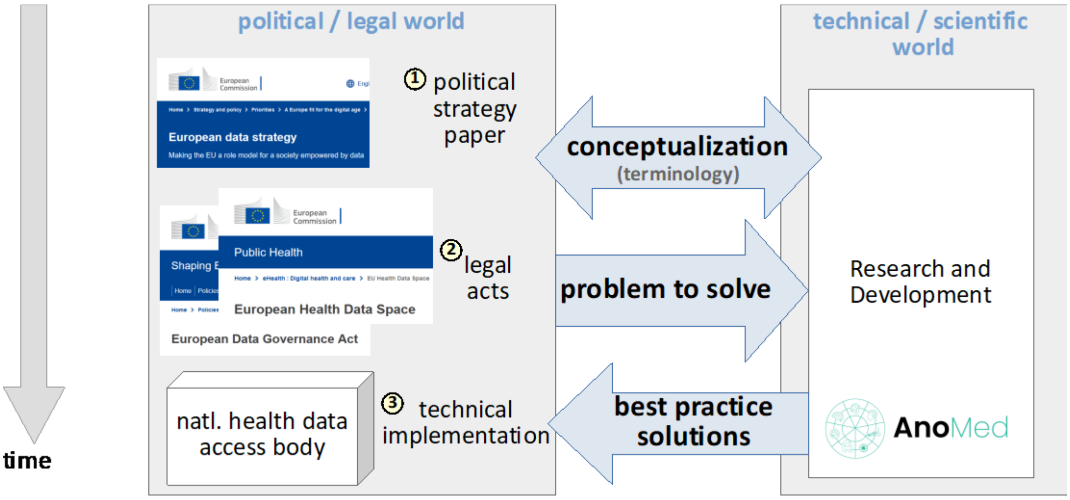

time
political / legal world
European Commission
European data strategy
Making the EU a role model for a society empowered by data
① political strategy paper
Public Health
European Health Data Space
Shaping E
European Data Governance Act
② legal acts
natl. health data access body
③ technical implementation
conceptualization
(terminology)
problem to solve
best practice solutions
technical / scientific world
Research and Development
AnoMed

Figure 3: Technology transfer in the context of the European Data Strategy.

Obviously, introducing new technology in society cannot be based solely on policy and legal decisions; much rather the decisions must be informed by the technical/scientific world. The term technology transfer is used in the sequel for everything that involves interaction between the policy/legal world and the technical/scientific world.

The possibly best-known aspect of technology transfer is the use of **best practice solutions that originate in the technical world** in the technical implementation. For this purpose, best practices have to be identified and matched with the implementation needs of the policy world. This is visualized by the blue bottom arrow in the figure.

A second kind of interaction is to inform players of the technical/scientific world of the technical artefacts that are used in legal texts. This can be seen as **requirements for the technical implementation**. From a technical perspective, this defines the problem space that has to be matched with the solution space. Since this interaction originates mostly originates in the legal world, ULD sometimes calls this type of interaction also **"legal transfer"**. Legal transfer is necessary to identify matching best practice solutions and to potentially identify

A third kind of interaction is desirable and should happen before the other two. It is that of **harmonizing the conceptualization** of the technical artefact between the technical/scientific world and the policy/legal world. While it falls short of a "common language", the languages used in the two disciplines should be compatible. In particular, legislators need to have realistic expectations when they describe technical artefacts in legal acts. Similarly, technical players need to understand the legal requirements for certain concepts such as "anonymization". The legal description of artefacts should accommodate the possibility of implementing it with state-of-the-art solutions.

The harmonizing of conceptualizations is also topic of other tasks of AnoMed (see for example Deliverables D4.9.1 and D4.9.2). The aspect covered by the terminology aims at fostering a better understanding of technology on the part of policy makers and legislators. While the concepts contained in the terminology are necessarily simplifying abstractions of the corresponding technical concepts, they are conceived with the objective of being more realistic, correct possible misconceptions, and avoid foreseeable misunderstandings.

To further illustrate the harmonization, Figure 4 shows examples for possible difference in the conceptualization of technical artefacts. In particular, the basic manner of reasoning is legal on the policy side while it is mathematical on the technical side. Further, the notion of anonymous is usually binary for policy makers (i.e., it falls under the GDPR: [y/n]) or a continuum expressing risk of re-identification (and for example characterized by epsilon in differential privacy). There may also be concepts such as “privacy budget” on the technical side that seemingly lack an equivalent in the legal conceptualization. This seems to follow for example from the fact that after anonymization, data is published on web servers without any further obligations.

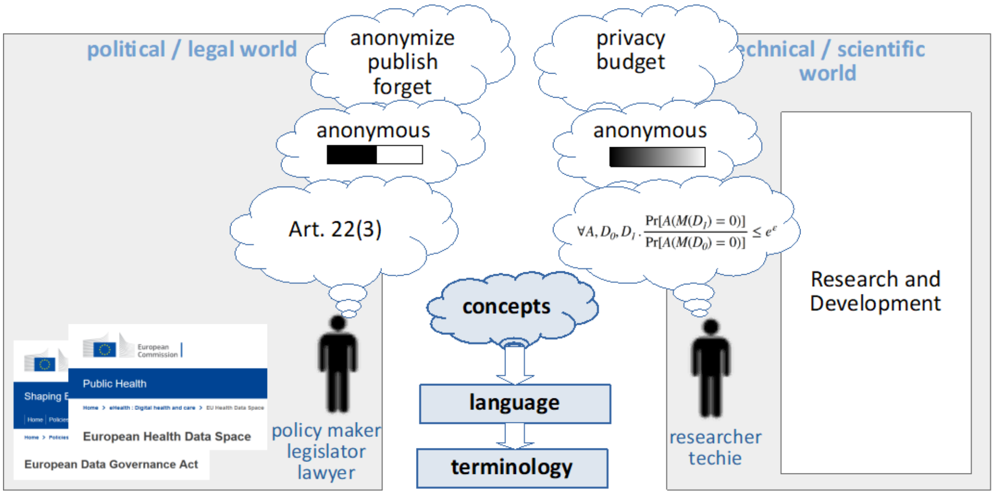


Figure 4: Examples of differences in conceptualization.

4 Analysis of Currently Conceptualizations

To reason about the harmonization of two conceptualizations, it is first necessary to understand these conceptualizations better. To find out what concepts are used, language can provide the necessary insight. In particular, the language uses terms to name concepts. How these terms are used, what is stated about them, and how they are related to other terms all gives insight about the concept behind it and its semantics. Language analysis, or the “reverse engineering” of text, thus permits to reason about whether two (possibly distinct) terms in the two domains refer to the same underlying concept (i.e., are synonyms) or whether the underlying concepts are only similar but semantically distinct (even if they may name the concept with the same term).

To gain insight in the conceptualization, substantial textual analysis has been conducted for both, scientific and legal texts.

To better understand the concepts related to “anonymization” on the **scientific side**, ULD analyzed a significant number of research articles about modern anonymization techniques. The almost weekly videoconferences between the project partners ULD and UzL (and later also UHH, conducted partly in UAP 4.9.5) were instrumental in this endeavor. Not only did ULD receive pointers to relevant scientific papers, but in addition important support to understand the concepts better and to discuss differences between legal and technical concepts.

To identify and understand the concepts on the **legal side**, the most relevant legal texts were analyzed by ULD. Namely, in addition to the already well-known *GDPR*, ULD studied the *Data Governance Act (DGA)*, the proposed European Health Data Space (EHDS), and monitored the evolving *Guidelines on Anonymization* by the *European Data Protection Board*[4] (EDPB, see also

[4] https://edpb.europa.eu/

Deliverable D4.9.2 which provides more details on the monitoring of the guidelines and some of the analysis work performed on the mentioned legal acts).

5 Harmonization Needs of Conceptualizations

The analysis of texts from both the technical/scientific, and the policy/legal side has identified some significant differences in how concepts are understood. To guarantee that policies/laws can be effectively technically implemented, a harmonization that addresses these differences seems to be beneficial. The following lists some of the major differences that were identified in the text analysis and how they can be addressed.

(i) A major difference concerns the **notion of anonymization and anonymous**.

In the **legal** world, data is either anonymous or not, i.e., anonymity is mostly understood as a **binary state**. Legally, data is anonymous if it is not personal data. Since only personal data are subject to the GDPR, anonymous data fall outside of its protection. The GDPR takes a risk-based approach to data protection where, proportional to the risk posed by the data processing, measures have to be implemented to protect data subjects. Consequently, the processing of anonymous data is understood as not posing any risk to data subjects. This becomes critical when considering that in the text analysis of recent legal acts, the outcome of "*anonymization*" of personal data seems to be consistently considered as being (successfully) anonymous data.

In the **technical** world in contrast, anonymization is not understood as a transformation that results in anonymous data. Much rather, it is understood that **every disclosure** of "anonymized" data **leaks information** about its data subjects. The amount of leakage is often measurable, for example by the value *epsilon* in *epsilon-differential privacy*. This epsilon is typically small but can never go down to zero. Thus, the technical concept of anonymization results in a **continuum** much rather than in two states [personal, anonymous].

To **address this difference**, it is important that in the legal world, the fallacy that *anonymization* results in *anonymous data* has to be made explicit. This erroneous implication may originate in the choice of terminology. The solution taken by the developed terminology is to refrain from using the misleading term *anonymization* all-together and use ***identity-reduction transformation*** instead. This newly proposed term is much more in line with the technical understanding, namely that the transformation merely reduces the identification potential but usually fails to eliminate it all together.

(ii) Another major difference in concepts is the implied believe that it is possible to **determine whether data is indeed anonymous**.

In **legal** texts, this determination is typically the responsibility of the controller of processing. There is ample legal analysis (based on Recital 26 GDPR) of how to determine whether data is personal (or anonymous). This also includes the upcoming *Guidelines on Anonymization*[5] by the EDPB. The existing legal analyses are limited to enumerating the factors that have to be considered in this

[5] See announcement in: EDPB, "EDPB Work Programme 2023/2024", adopted 13 February 2023, https://edpb.europa.eu/system/files/2023-02/edpb_work_programme_2023-2024_en.pdf.

decision, such as relate-to/singling-out, identifiable person/linkage, and inference and the means reasonable likely to be available to an "attacker". This falls short of specifying how to make the determination. In particular, there remains an unanswered question of threshold. For example, which degree of "unlinkability", or which degree of "unlikelihood", are sufficient to determine that data are indeed anonymous. The determination thus depends on the risk tolerance of controllers.

In **technical** texts, it is always clear that, on a continuous scale of information leakage, such a yes/no determination cannot be made. Also, there is clearly no technical method to determine whether data is indeed anonymous.

To **address this difference**, the terminology contains **four possible outcomes of *identity-reduction transformations*** which capture the uncertainty of the underlying determination. In particular, the outcomes can be:

- *Basic pseudonymous data* where direct identification is no longer possible;
- *Advanced pseudonymous data* where obvious re-identification is prevented, but which are likely still personal;
- *Supposedly anonymous data*[6] that is unlikely to be re-identifiable but where the controller cannot exclude such a possibility;
- *Successfully anonymous data* where the controller has certainty that re-identification is not possible.

In consequence, the terminology renders the subjective character of determining whether data is anonymous explicit. It captures the usually non-zero residual risk of re-identification[7]. The event of unexpected re-identification (legally considered a data breach) is also well accommodated in the proposed terminology. In contrast, with the binary notion of anonymity, it is difficult to explain within the logical model how data that was anonymous and outside the scope of the GDPR can suddenly become personal again and fall inside the GDPR (and do so as a violation).

(iii) Another major difference in concepts lies in the **diversification of the concept of anonymization**.

In **legal** texts, there is the **single concept of *anonymization*** that is not further diversified and considered to result in anonymous data. The closest to a diversification is the concept of *pseudonymous* that is used alongside of anonymous. While in the texts, *anonymization* never results in *pseudonymous* data, clearly the pseudonymization of data is understood as a manner of reducing that identification potential of data.

In **technical** texts, in contrast, a **wide variety of *identity-reduction transformations*** is evident. They vary widely in their scope, the guarantees they provide, the assumptions that are necessary for a successful use, etc. Some technical texts also strongly discourage the use of certain (families of) identity-reduction transformations (such as K-Anonymity) for the legal concept of *anonymization*, since such transformations are unable to provide sufficient guarantees of privacy. The technical literature is full of demonstrations how presumed anonymous data created with such transformations can be re-identified.

[6] See the equivalent concept of *presumed anonymous data* previously proposed by the author of this deliverable within the EU-funded PANELFIT project, in the report *Towards a Better Understanding of Identification, Pseudonymization, and Anonymization*, definition of term in section 5.3, https://uldsh.de/pseudoanon/, last visited 22/2/2024.

[7] Note that also the European Health Data Space proposal acknowledges the residual risk of re-identification in the 2nd sentence of its Recital 64 but then fails to address the case of an unexpected re-identification actually happening.

To **address this difference**, in a first step, a **taxonomy** consisting of three major kinds (and six more detailed kinds) **of identity-reduction transformations** is part of the terminology. This is meant to replace the single concept of *anonymization*. One of the key differences between such transformations is their scope of protection, ranging from part of a data set, over the whole data set, up to multiple disclosures (i.e., data sets). For example K-Anonymity only protects against re-identification based on the part of the data set containing "quasi identifier"; it fails to provide any protection against re-identification using other parts of the data set, for example an exceptional attribute value[8].

Consequently, in comparison to the single concept of *anonymization*, the proposed taxonomy significantly diversifies the concept and makes it (even graphically) explicit, that identity-reduction transformations have their limitations. The diversification provides more than just one word to talk about "anonymization" and is hoped to open the door to reasoning that is closer to the technical reality.

At the time of writing, it is considered to make further steps to address this issue in a separate terminology of *guarantees* provided by *identity-reducing transformations*.

(iv) Another major difference in concepts lies in the apparent **scope of "anonymization"**.

In **legal** texts, it seems that the prototypical thinking behind "anonymization" **considers only a single data set**. The combination of multiple "anonymized" data sets does not seem to be considered. A reason for this could be that the focus of the GDPR is a single processing activity of (mostly) a single controller and that the combination of "anonymized" data from multiple controllers (and processing activities) is considered to be another processing activity. The latter may be considered to be illegal by the law[9], but could nevertheless result in re-identified data.

In **technical** texts, it is **common to consider multiple disclosures** of related data about the same data subjects. More precisely, this is often captured by the "*privacy budget*" which erodes with every disclosure. Many technical texts propose identity-reducing transformations designed for multiple disclosures (much rather a single data set). They can capture how additional disclosures erode the privacy budget[10]. Well established methods can thus be used to control multiple disclosures by the same controller; it is more difficult how to manage multiple disclosures by different controllers (who may not know about each other). Since every disclosure erodes the privacy budget, the **risk of re-identification also augments with every disclosure**.

To better understand how multiple disclosures can lead to re-identification, consider that every disclosure leaks some information. When enough such information comes together, it is then possible to reconstruct information about individual data subjects. The U.S. Bureau of the Census has demonstrated that for their own 2010 census[11]. Here, a large number of disclosed statistics have been fed into a very large equation system that could be solved with significant computational effort resulting in reconstructed information about an unexpectedly large percentage of data subjects.

[8] Examples for exceptional attributes include an exceptionally tall or old person, i.e., properties that can easily be unique and thus highly identifying in a given set of candidates.

[9] For example, it may be difficult to find a valid legal basis according to Article 6 GDPR for an attempt to re-identify "anonymized" data originating from multiple distinct processing activities.

[10] The erosion can be captured thanks to the composability of the method).

[11] *Abowd, J.M., Adams, T., Ashmead, R., Darais, D., Dey, S., Garfinkel, S.L., Goldschlag, N., Kifer, D., Leclerc, P., Lew, E., Moore, S., Rodr'iguez, R.A., Tadros, R.N., & Vilhuber, L. (2023). The 2010 Census Confidentiality Protections Failed, Here's How and Why. ArXiv, abs/2312.11283,* https://arxiv.org/abs/2312.11283, *last visited 22/2/2024.*

Similarly, if anonymized data is protected by noise injection (as is the case with differential privacy), multiple disclosures (i.e., noise realizations) can be used to average out the noise.

To **address this difference**, in a first step, the terminology makes a distinction of transformations whose scope is a single data set and so called **disclosure control**. Disclosure control stands for transformations whose scope comprise multiple disclosures of "anonymized" data. The terminology attempts to make **multiple disclosures graphically visible**. While being a modest step, it at least raises the awareness of multiple disclosures and that these affect the risk of re-identification. The issue may be further addressed in the distinct terminology concerned with guarantees.

6 Requirements for the Terminologies

This section briefly describes requirements for the developed terminologies. In the effort to harmonize conceptualizations between the technical and the legal worlds, the terminology is the "message" that is sent from the technical to the legal side.

The comparison of technical and legal conceptualization in the previous section has uncovered several differences where the legal concepts may not be technically realistic. The harmonization effort therefore must send "corrective messages" from the technical to the legal side.

Consequently the main audience of the harmonization effort are policy makers and legislators. Since they are typically not mathematically versed and usually very busy, communication is only effective if appositely targeted at that kind of audience.

The targeting leads to two major requirements:

i. The terminology (i.e., message) has to be simple, short, and attractive
ii. The terminology needs to be compatible with other terminologies used by the targeted policy makers.

(i) To keep the terminology simple, it has to refrain from using overly technical terms or methods of communication. It is desirable to express as much as possible in a "visible" (graphical) manner much rather than require reading. Showing relationships though graphical position and create connections though color coding can be helpful. The number of introduced terms must be as reduced as possible for the necessary message. The terminology should be as attractive and practically useful as possible. Again, graphical expression and the form-factor of handouts helps here.

(ii) The targeted policy makers already use related terminologies. To foster acceptance and adoption of the proposed terminology, its terms have to be as compatible as possible with the other terminologies. For this purpose, the following (implicit) terminologies have been studied and have provided input in the choice of terms:

- The GDPR,
- The DGA and EHDS,
- The evolving EDPB guidelines on anonymization that were monitored (see D4.9.2).
- The evolving EDPB guidelines on pseudonymization that were monitored (see D4.9.2).

7 Terminologies developed for AnoMed

As a contribution to technology transfer and based on the extensive analysis of texts and associated terminologies described above, two terminologies have been developed in AnoMed that are described in the following and are included in (almost) full size in the Appendix.

7.1 Identity-Reduction Terminology

While the actual terminology can be found in the appendix, the following shows it in reduced size and points out its major properties.

The terminology consists of four sheets. One of which is a figure, two are tables, and one is a textual glossary of the terms used in the other pages.

The relation between the first three pages is establishes by color coding. In particular, two color schemes are used: one for the scope of identity-reduction transformations, and one for the possible outcomes.

Figure 5 shows the first page of the terminology.

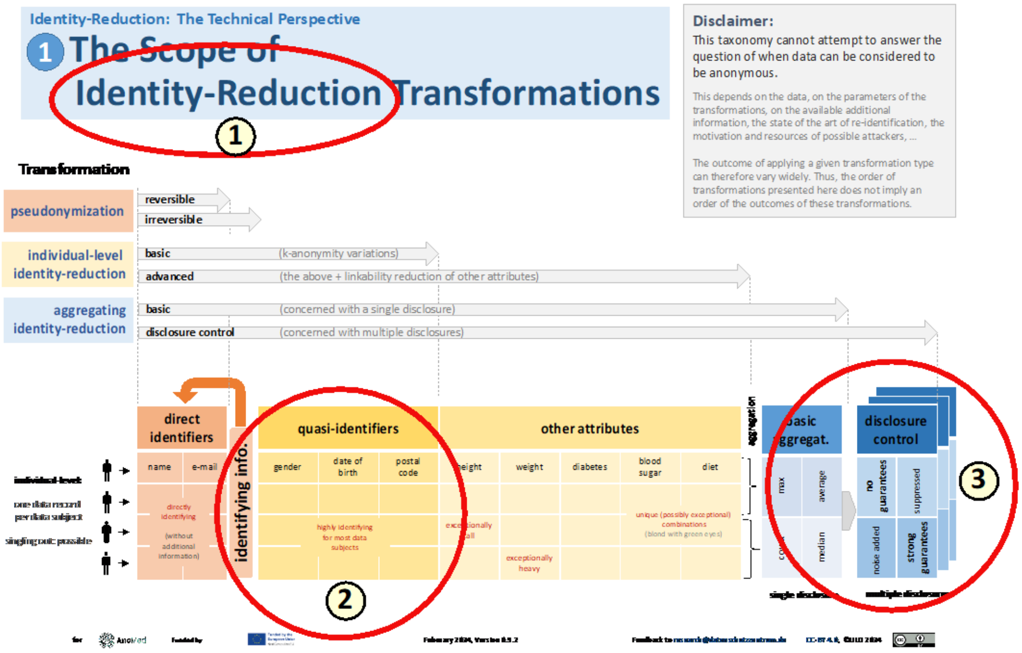
The most central characteristic (1) is in its title and is the use of the term “identity-reduction” much rather than “anonymization”.

At the bottom of the page, in reddish and yellow, a prototypical tabular data set is shown. In it, three different parts are distinguished: direct identifying attributes, quasi-identifiers, and other attributes. The prototypical illustration of data is then extended towards the right to show derived aggregated data (such as statics) and the widening of the scope from a single “anonymized” data set (i.e., disclosure) to multiple disclosures. The cells of the prototypical data contain examples to ease understanding of the structure and the terms that are used. This could be seen as definition of the used terms by example. Alternatively, the same terms are defined by text in the glossary of page 4.

The prototypical representation of data in the bottom part of the figure are then used to define different scopes of identity-reduction transformations. The scope is then expressed by the length of the shown arrows. The scope underlies the proposed taxonomy of identity-reduction transformation. For simplicity, only three major scopes that are close to policy makers’ current usage of terms are proposed.

By showing that a the scope of a transformation can be also only a subset of the data (2), it becomes visually evident, that the protection against re-identification by such transformations is limited. In particular, identification based on parts of data that lie outside of this limited scope can obviously not be prevented by such transformations.

The figure also attempts to widen the awareness of policy makers by visualizing multiple disclosures (3). In particular, multiple disclosures are represented graphically as an easy to understand multitude of “anonymized” data.



Disclaimer:
This taxonomy cannot attempt to answer the question of when data can be considered to be anonymous.

This depends on the data, on the parameters of the transformations, on the available additional information, the state of the art of re-identification, the motivation and resources of possible attackers, …

The outcome of applying a given transformation type can therefore vary widely. Thus, the order of transformations presented here does not imply an order of the outcomes of these transformations.

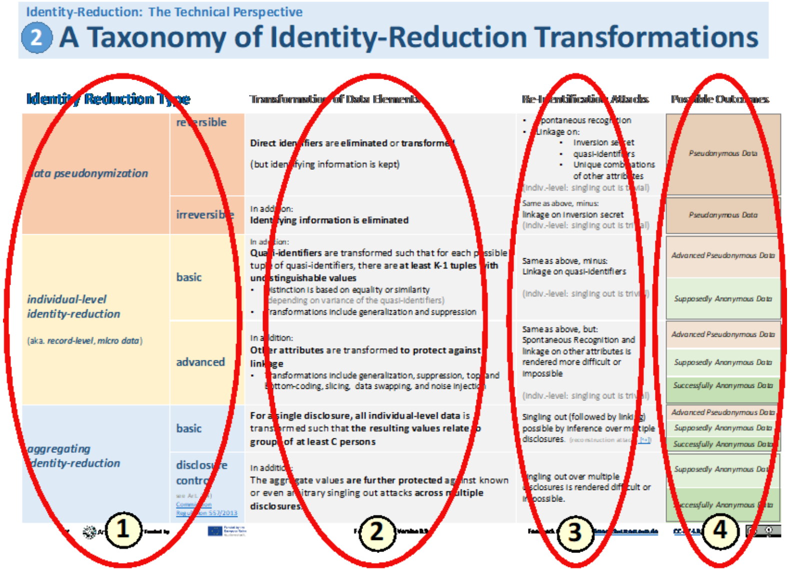
Figure 5: First page of Identity-Reduction Terminology.

Figure 6 shows the second page of the terminology. It takes up the different scopes of the first page (1) and provides a textual definition (2) (much rather than a definition by example of the first page). By using the scope as a taxonomy, it introduces *terms* (1) to refer to different kinds of identity-reduction transformations. Providing six terms in place of the previously single term of “anonymization”, it promises to support more diversified discussions that are also able to capture limitations of “anonymization” techniques.

To further support the perception of such limitations, the table shows that all transformations are still subject to re-identification attacks (3). In other words, none of the technically known transformations provides absolute guarantees that re-identification is not possible. Even in the case of differential privacy, re-identification may be possible when the privacy budget is sufficiently eroded; and such erosion may be caused by multiple controllers who do not know about each other.

The table makes it evident (3) that widening the scope of identity-reducing transformations eliminates whole classes of re-identification vulnerabilities.

To express the difficulty of determining the outcome of identity-reducing transformations, the table shows (4) how each transformation can result in different outcomes. These outcomes are color coded and link to page 3 of the terminology.



Identity-Reduction: The Technical Perspective

2 A Taxonomy of Identity-Reduction Transformations

Identity Reduction Type		Transformation of Data Elements	Re-Identification Attacks	Possible Outcomes
data pseudonymization	reversible	Direct identifiers are eliminated or transformed (but identifying information is kept)	• Spontaneous recognition • Linkage on: • Inversion secret • quasi-identifiers • Unique combinations of other attributes (indiv.-level: singling out is trivial)	*Pseudonymous Data*
	irreversible	In addition: Identifying information is eliminated	Same as above, minus: linkage on inversion secret (indiv.-level: singling out is trivial)	*Pseudonymous Data*
individual-level identity-reduction (aka. *record-level*, *micro data*)	basic	In addition: Quasi-identifiers are transformed such that for each possible tuple of quasi-identifiers, there are at least K-1 tuples with undistinguishable values • Distinction is based on equality or similarity (depending on variance of the quasi-identifiers) • Transformations include generalization and suppression	Same as above, minus: Linkage on quasi-identifiers (indiv.-level: singling out is trivial)	*Advanced Pseudonymous Data* *Supposedly Anonymous Data*
	advanced	In addition: Other attributes are transformed to protect against linkage • Transformations include generalization, suppression, top and bottom-coding, slicing, data swapping, and noise injection	Same as above, but: Spontaneous Recognition and linkage on other attributes is rendered more difficult or impossible (indiv.-level: singling out is trivial)	*Advanced Pseudonymous Data* *Supposedly Anonymous Data* *Successfully Anonymous Data*
aggregating identity-reduction	basic	For a single disclosure, all individual-level data is transformed such that the resulting values relate to groups of at least C persons	Singling out (followed by linking) possible by inference over multiple disclosures. (reconstruction attacks [*+])	*Advanced Pseudonymous Data* *Supposedly Anonymous Data* *Successfully Anonymous Data*
	disclosure control see Art. 2(i) Commission Regulation 557/2013	In addition: The aggregate values are further protected against known or even arbitrary singling out attacks across multiple disclosures.	Singling out over multiple disclosures is rendered difficult or impossible.	*Supposedly Anonymous Data* *Successfully Anonymous Data*

Figure 6: Second page of Identity-Reduction Terminology.

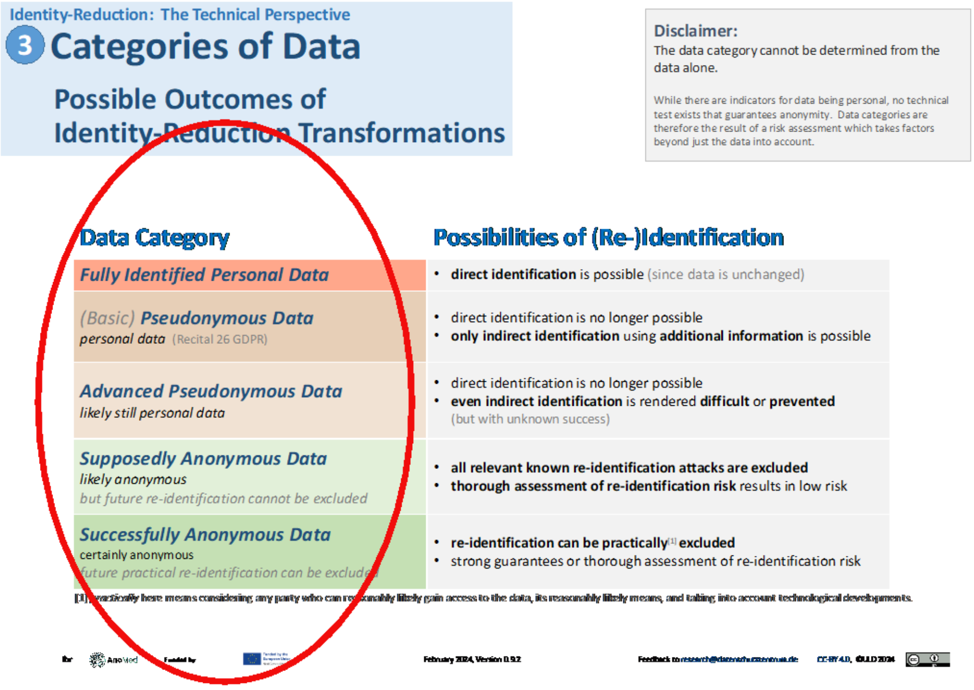
Figure 7 shows the third page of the terminology on possible categories of data (highlighted in a red ellipses). *Fully identified personal* data is usually considered to be the input to identity-reduction transformations; the other four are possible outcomes.

The taxonomy of outcomes uses five categories instead of the commonly used three in the legal world: (fully identified) personal data, pseudonymous (still personal) data, and anonymous data. It thus tried to keep it simple for policy makers by staying close to their current conceptualization and pushing the boundary only minimally in a hopefully easy step.

The five terms are also chosen to be mostly compatible with terminologies already used by policy makers including the GDPR, the DGA and EHDS, as well as the evolving *Guidelines on Anonymization* by the EDPB.

The textual description of possible re-identification (gray) serves to further establish the link to the legal concepts in the GDPR.

The color coding from red (a warning color) to dark green (an ok color) attempts to express a level of risk of re-identification. The color and the arrangement of terms in the table establishes a clear order relation between the terms. Again, the order is visual and thus doesn't have to be explained by text.



Identity-Reduction: The Technical Perspective

3 Categories of Data

Possible Outcomes of Identity-Reduction Transformations

Disclaimer:
The data category cannot be determined from the data alone.

While there are indicators for data being personal, no technical test exists that guarantees anonymity. Data categories are therefore the result of a risk assessment which takes factors beyond just the data into account.

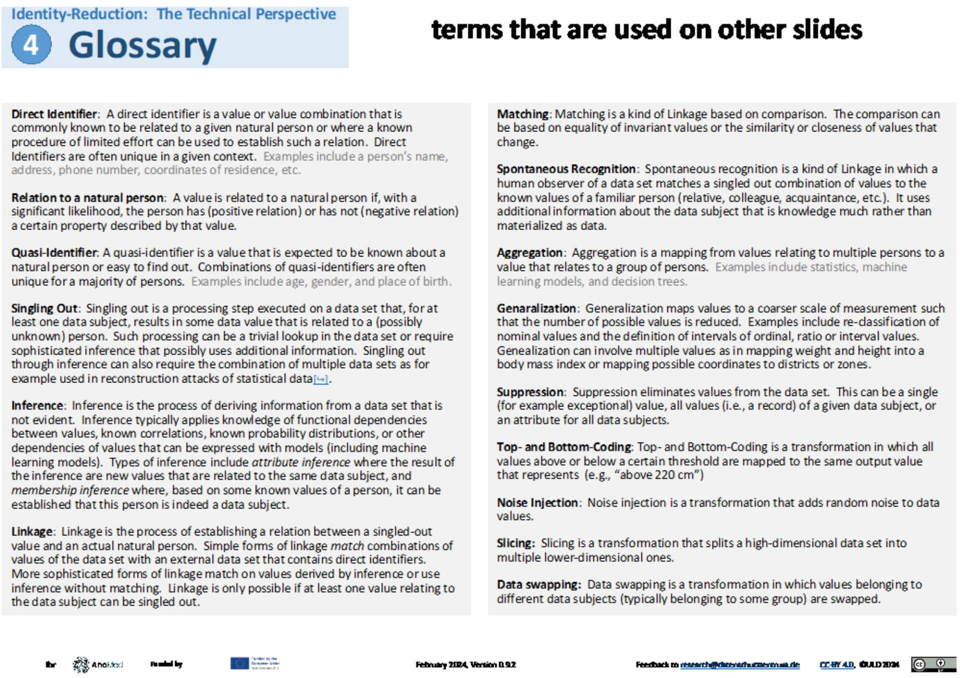
Data Category	Possibilities of (Re-)Identification
Fully Identified Personal Data	• **direct identification** is possible (since data is unchanged)
(Basic) Pseudonymous Data *personal data* (Recital 26 GDPR)	• direct identification is no longer possible • **only indirect identification** using **additional information** is possible
Advanced Pseudonymous Data *likely still personal data*	• direct identification is no longer possible • **even indirect identification** is rendered **difficult** or **prevented** (but with unknown success)
Supposedly Anonymous Data *likely anonymous* *but future re-identification cannot be excluded*	• **all relevant known re-identification attacks are excluded** • **thorough assessment of re-identification risk** results in low risk
Successfully Anonymous Data certainly anonymous *future practical re-identification can be excluded*	• **re-identification can be practically[1] excluded** • strong guarantees or thorough assessment of re-identification risk

[1] practically here means considering any party who can reasonably likely gain access to the data, its reasonably likely means, and taking into account technological developments.

for AnoMed Funded by February 2024, Version 0.9.2 Feedback to research@datenschutzzentrum.de CC-BY 4.0, ©ULD 2024

Figure 7: Third page of Identity-Reduction Terminology.

Figure 8 shows the fourth page of the terminology. It is a classical textual glossary. While it will likely not be very attractive to policy makers, it can be useful for detail understanding and is a useful way of filling the otherwise empty back side in a two sheet handout. The glossary can also be used for other purposes, such as part of reports and essays in the domain of anonymization.



Identity-Reduction: The Technical Perspective

4 Glossary

terms that are used on other slides

Direct Identifier: A direct identifier is a value or value combination that is commonly known to be related to a given natural person or where a known procedure of limited effort can be used to establish such a relation. Direct Identifiers are often unique in a given context. Examples include a person's name, address, phone number, coordinates of residence, etc.

Relation to a natural person: A value is related to a natural person if, with a significant likelihood, the person has (positive relation) or has not (negative relation) a certain property described by that value.

Quasi-Identifier: A quasi-identifier is a value that is expected to be known about a natural person or easy to find out. Combinations of quasi-identifiers are often unique for a majority of persons. Examples include age, gender, and place of birth.

Singling Out: Singling out is a processing step executed on a data set that, for at least one data subject, results in some data value that is related to a (possibly unknown) person. Such processing can be a trivial lookup in the data set or require sophisticated inference that possibly uses additional information. Singling out through inference can also require the combination of multiple data sets as for example used in reconstruction attacks of statistical data[↪].

Inference: Inference is the process of deriving information from a data set that is not evident. Inference typically applies knowledge of functional dependencies between values, known correlations, known probability distributions, or other dependencies of values that can be expressed with models (including machine learning models). Types of inference include *attribute inference* where the result of the inference are new values that are related to the same data subject, and *membership inference* where, based on some known values of a person, it can be established that this person is indeed a data subject.

Linkage: Linkage is the process of establishing a relation between a singled-out value and an actual natural person. Simple forms of linkage *match* combinations of values of the data set with an external data set that contains direct identifiers. More sophisticated forms of linkage match on values derived by inference or use inference without matching. Linkage is only possible if at least one value relating to the data subject can be singled out.

Matching: Matching is a kind of Linkage based on comparison. The comparison can be based on equality of invariant values or the similarity or closeness of values that change.

Spontaneous Recognition: Spontaneous recognition is a kind of Linkage in which a human observer of a data set matches a singled out combination of values to the known values of a familiar person (relative, colleague, acquaintance, etc.). It uses additional information about the data subject that is knowledge much rather than materialized as data.

Aggregation: Aggregation is a mapping from values relating to multiple persons to a value that relates to a group of persons. Examples include statistics, machine learning models, and decision trees.

Genaralization: Generalization maps values to a coarser scale of measurement such that the number of possible values is reduced. Examples include re-classification of nominal values and the definition of intervals of ordinal, ratio or interval values. Genealization can involve multiple values as in mapping weight and height into a body mass index or mapping possible coordinates to districts or zones.

Suppression: Suppression eliminates values from the data set. This can be a single (for example exceptional) value, all values (i.e., a record) of a given data subject, or an attribute for all data subjects.

Top- and Bottom-Coding: Top- and Bottom-Coding is a transformation in which all values above or below a certain threshold are mapped to the same output value that represents (e.g., "above 220 cm")

Noise Injection: Noise injection is a transformation that adds random noise to data values.

Slicing: Slicing is a transformation that splits a high-dimensional data set into multiple lower-dimensional ones.

Data swapping: Data swapping is a transformation in which values belonging to different data subjects (typically belonging to some group) are swapped.

for AnoMed Funded by February 2024, Version 0.9.2 Feedback to research@datenschutzzentrum.de CC-BY 4.0, ©ULD 2024

Figure 8: Fourth page of Identity-Reduction Terminology.

7.2 *Pseudonymization* Terminology

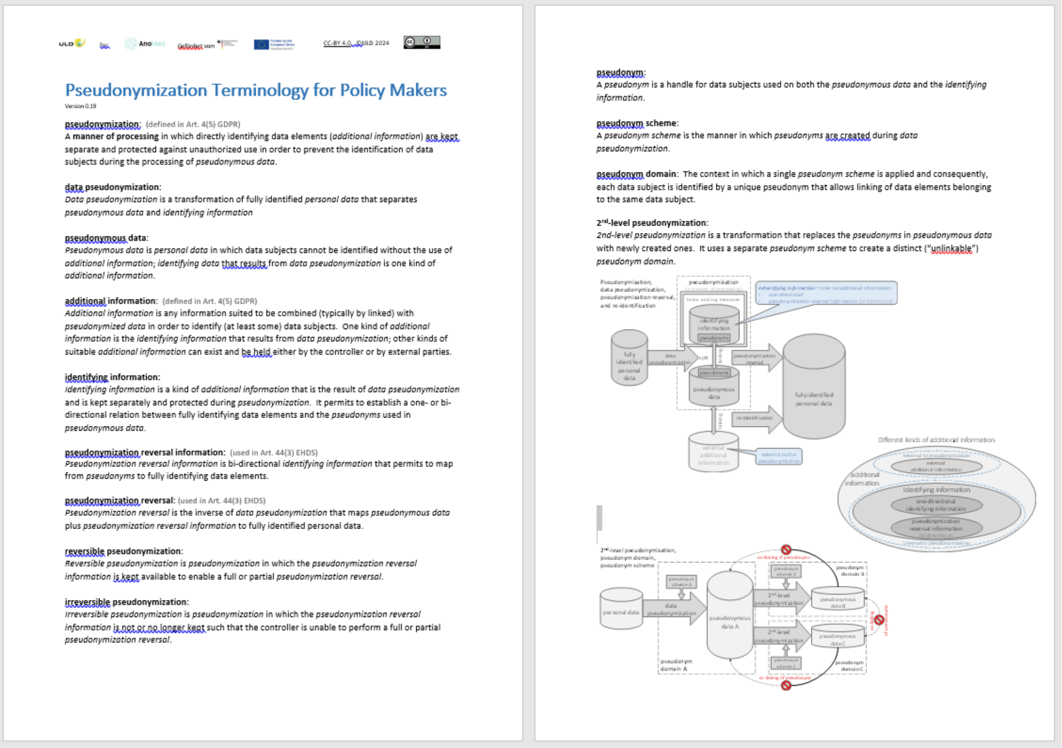
The second terminology on pseudonymization takes a more classical approach by providing a textual definition for every term. The attractiveness to policy-makers is hoped to come from the visualization of all terms in three figures. These figures may be the entry point that in many cases may provide a graphical definition of terms; the textual definition may be consumed in a second step to render the concepts more precise and unambiguous.

The terminology is based on a careful selection of terms form a surprising number of alternatives. Also care was taken to stay compatible with the terminology already used or proposed.

The terminology provides a complete set of mutually compatible terms that can support the writing of legal texts, for example by legislators concerned with European data spaces. The terms "pseudonym domain" and "2nd level pseudonymization" were included to make policy makers aware of the full set of technical possibilities in the hope to still have effect on the final version of the EHDS[12] or the subsequent implementation of data access bodies and their concepts for data flows, processes technical and organizational measures.

The terminology was developed in a long and a short version. The longer version provides more detail and is suited to provide clarity to a more technically-oriented audience. The short version is a subset of the long version and is more geared to policy makers. It fits on a single sheet of a double-sided handout.

Figure 9 shows the short version of the terminology.



CC-BY 4.0, ©ULD 2024

Pseudonymization Terminology for Policy Makers

Version 0.19

pseudonymization: (defined in Art. 4(5) GDPR)
A manner of processing in which directly identifying data elements (*additional information*) are kept separate and protected against unauthorized use in order to prevent the identification of data subjects during the processing of *pseudonymous data*.

data pseudonymization:
Data pseudonymization is a transformation of fully identified *personal data* that separates *pseudonymous data* and *identifying information*

pseudonymous data:
Pseudonymous data is *personal data* in which data subjects cannot be identified without the use of *additional information*; *identifying data* that results from *data pseudonymization* is one kind of *additional information*.

additional information: (defined in Art. 4(5) GDPR)
Additional information is any information suited to be combined (typically by linked) with *pseudonymized data* in order to identify (at least some) data subjects. One kind of *additional information* is the *identifying information* that results from *data pseudonymization*; other kinds of suitable *additional information* can exist and be held either by the controller or by external parties.

identifying information:
Identifying information is a kind of *additional information* that is the result of *data pseudonymization* and is kept separately and protected during *pseudonymization*. It permits to establish a one- or bi-directional relation between fully identifying data elements and the *pseudonyms* used in *pseudonymous data*.

pseudonymization reversal information: (used in Art. 44(3) EHDS)
Pseudonymization reversal information is bi-directional *identifying information* that permits to map from *pseudonyms* to fully identifying data elements.

pseudonymization reversal: (used in Art. 44(3) EHDS)
Pseudonymization reversal is the inverse of *data pseudonymization* that maps *pseudonymous data* plus *pseudonymization reversal information* to fully identified personal data.

reversible pseudonymization:
Reversible pseudonymization is *pseudonymization* in which the *pseudonymization reversal information* is kept available to enable a full or partial *pseudonymization reversal*.

irreversible pseudonymization:
Irreversible pseudonymization is *pseudonymization* in which the *pseudonymization reversal information* is not or no longer kept such that the controller is unable to perform a full or partial *pseudonymization reversal*.

pseudonym:
A *pseudonym* is a handle for data subjects used on both the *pseudonymous data* and the *identifying information*.

pseudonym scheme:
A *pseudonym scheme* is the manner in which *pseudonyms* are created during *data pseudonymization*.

pseudonym domain: The context in which a single *pseudonym scheme* is applied and consequently, each data subject is identified by a unique pseudonym that allows linking of data elements belonging to the same data subject.

2nd-level pseudonymization:
2nd-level pseudonymization is a transformation that replaces the *pseudonyms* in *pseudonymous data* with newly created ones. It uses a separate *pseudonym scheme* to create a distinct ("unlinkable") *pseudonym domain*.

Pseudonymization, data pseudonymization, pseudonymization reversal, and re-identification
pseudonymization
(a manner of processing)
techn. and org. measures
identifying information
pseudonyms
identifying information = internal additional information:
one-directional
pseudonymization reversal information (bi-directional)
fully identified personal data
data pseudonymization
split
linking
pseudonymization reversal
pseudonyms
pseudonymous data
fully identified personal data
re-identification
linking
external additional information
external to the pseudonymization
Different kinds of additional information
external to pseudonymization
external additional information
additional information
identifying information
one-directional identifying information
pseudonymization reversal information
(bi-directional)
internal to pseudonymization
2nd-level pseudonymization, pseudonym domain, pseudonym scheme
no linking of pseudonyms
pseudonym scheme A
pseudonym scheme B
pseudonym domain B
personal data
data pseudonymization
pseudonymous data A
2nd-level pseudonymization
pseudonymous data B
no linking of pseudonyms
2nd-level pseudonymization
pseudonymous data C
pseudonym scheme C
pseudonym domain A
pseudonym domain C
no linking of pseudonyms

Figure 9: The developed pseudonymization terminology (short version).

[12] The most recent version is a proposal by the EC that is the basis for negotiations in the trilogy and will necessarily change before being approved and published in the official journal.

8 Dissemination of the Terminologies

The terminologies that were developed in AnoMed were disseminated both internally in the project and externally.

Internally, regular video conferences between UzL and ULD already shaped the creation of the terminologies. They were further used to collect feedback and validate that they are congruent with the technical reality. Once that has happened, the terminology and the reasoning behind them was disseminated to a wider circle of more project partners in the form of the third module of the work shop (see D4.9.1 for details). The work shop aimed at motivating and enabling partners to use and further disseminate the terminologies.

The real impact of the terminologies is hoped to materialize in the context of technology transfer, however. This requires dissemination to external stakeholders, including policy makers, decision makers and practitioners in industry, research and academia.

The relevant recipients of the messages built into the terminologies are:

- The drafting team for the *Guidelines on Anonymization* of the EDPB, and
- Various kinds of policy makers who will evolve the current proposal of the European Health Data Space and in the future other data spaces.

This kind of audience cannot be easily reached with dissemination approaches typical in science such as publication of scientific papers or presentations at scientific conferences.

Since the target audience is relatively small and ULD has direct contact to some experts of that audience, a strategy of word of mouth was chosen. In particular, in a first step, ULD reaches out to a selected set of persons (so called “champions”) in the hope that they like the product and will disseminate it further to other experts who cannot be directly reached by AnoMed partners.

To support this indirect dissemination, two supportive actions have been taken:

- Selected experts were asked for feedback on a preliminary version of the terminologies in order to make them participate in its creation,
- the final product is licensed by a creative commons license that permits 3rd parties to further disseminate the terminology, and
- ULD has created a specific e-mail address that is listed on the handouts and permits to harvest feedback from users.

The dissemination of the terminologies is also hoped to be supported by their unique “market position”. With the advent of data spaces in Europe, understanding pseudonymization and anonymization have significantly gained importance. Most material on these topics is relatively complex and either strongly technical or legal. For example, the evolving EDPB *Guidelines on Anonymization* are long and consist of legal analysis that is not easy to grasp for non-lawyers. Considering that the developed terminologies are far shorter, more structured, and more technical, its creators hope that they have positioned them in a significant niche of the market where they currently seem to lack direct competition.

Following the above strategy, ULD has started with its dissemination effort including the following:

- Seven champions were identified to be asked to provide feedback on a preliminary version of the terminology.

- Three of them (two of whom working for national Data Protection Supervisory Authorities, DPAs) provided feedback.
- The feedback of one of these (from a major DPA) provided feedback that directly led to a new “fixed” version.
- Official approval of ULD acting as an official DPA on different versions was sought in order to improve perception of the terminologies by other DPAs.

In addition, both terminologies were printed as handouts to be distributed at MWC Barcelona 2024[13]. In particular, Marit Hansen, the head of the ULD distributed the handouts at the “*Digital Health & Wellness Summit - Data and Digital: Data Protection Challenges by Using the European Health Data Space (EHDS)*”[14]. It is hoped to reach some policy makers around the EHDS.

The session is describe as follows: “*The Catalan Data Protection Authority organizes this panel to analyze European Health Data Space (EHDS) from a data protection perspective. The session will offer an input on the need for proper pseudonymisation, how to achieve it, and the remaining data protection risks. The conference will also explore the GDPR compliant implementation, especially data protection by design and by default.*” It seems to be a very good match for terminologies on pseudonymization and anonymization.

9 Conclusions

Two short terminologies on pseudonymization and anonymization were developed in AnoMed. They pursue the larger objective of bridging the gap between science and policy. They attempt to enable AnoMed to contribute to the European strategy of large-scale reuse of data that promises important benefits for our society.

In the case of reuse of personal data, these important benefits are contrasted by the risk of unintended disclosure of personal data. This materialized when supposedly anonymous data can unexpectedly be re-identified. Consequently, to harvest the benefits at minimal risk, policy decisions have to be well-informed. This deliverable has described this process by providing some background on technology transfer.

To contribute to informed policy decisions, AnoMed has analyzed the conceptualization that is implied by both, legal/policy and technical/scientific, texts. This analysis resulted in the identification of certain mismatches between the description of technical artefacts in legal/policy texts and technical reality.

Two terminologies have been developed that policy makers can use and that convey a more harmonized conceptualization that is compatible with the technical understanding. Using these more realistic concepts for describing the policy decisions, it becomes more likely that the outcome can find a successful technical implementation.

The deliverable presents the terminologies and how they incorporate measures to harmonize the conceptualization between the legal/policy and technical/scientific worlds. It also reports on initial dissemination efforts that are crucial to create an impact.

[13] https://www.mwcbarcelona.com/, last visited 26/2/2024.

[14] https://www.mwcbarcelona.com/agenda/sessions/4376-digital-health-wellness-summit-data-and-digital-data-protection-challenges-by-using-the-european-health-data-space-ehds, last visited 26/2/2024.

10 Appendix

The Appendix contains the two developed terminologies in full size.

10.1 Identity-Reduction Terminology

The following includes the developed Identity-Reduction Terminology in (almost) original size. It consists of four A4 pages that can easily be printed front and back on two handout sheets. The attached version represents the status as of February 2024. As an update and further development is desired, the terminology is published under a creative commons license. Future feedback may result in updated or consolidated versions.

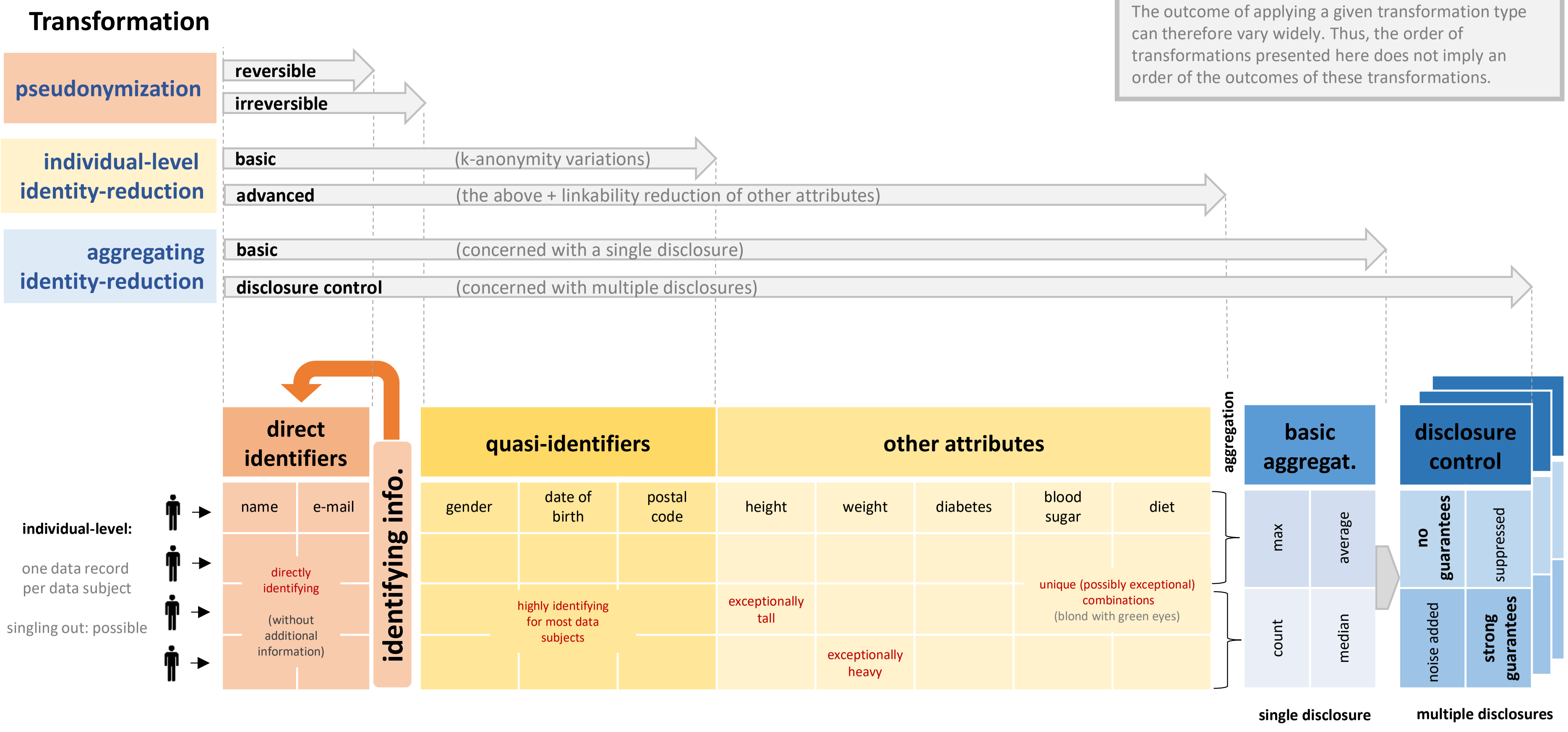
Identity-Reduction: The Technical Perspective

1 The Scope of Identity-Reduction Transformations

Disclaimer:
This taxonomy cannot attempt to answer the question of when data can be considered to be anonymous.

This depends on the data, on the parameters of the transformations, on the available additional information, the state of the art of re-identification, the motivation and resources of possible attackers, …

The outcome of applying a given transformation type can therefore vary widely. Thus, the order of transformations presented here does not imply an order of the outcomes of these transformations.



ULD for AnoMed Funded by Funded by the European Union NextGenerationEU

February 2024, Version 0.9.2

Feedback to research@datenschutzzentrum.de

Identity-Reduction: The Technical Perspective

2 A Taxonomy of Identity-Reduction Transformations

Identity Reduction Type		Transformation of Data Elements	Re-Identification Attacks	Possible Outcomes
data pseudonymization	**reversible**	**Direct identifiers** are **eliminated** or **transformed** (but identifying information is kept)	• Spontaneous recognition • Linkage on: • Inversion secret • quasi-identifiers • Unique combinations of other attributes (indiv.-level: singling out is trivial)	*Pseudonymous Data*
	irreversible	In addition: **Identifying information is eliminated**	Same as above, minus: linkage on inversion secret (indiv.-level: singling out is trivial)	*Pseudonymous Data*
individual-level identity-reduction (aka. ***record-level***, ***micro data***)	**basic**	In addition: **Quasi-identifiers** are transformed such that for each possible tuple of quasi-identifiers, there are **at least K-1 tuples with undistinguishable values** • Distinction is based on equality or similarity (depending on variance of the quasi-identifiers) • Transformations include generalization and suppression	Same as above, minus: Linkage on quasi-identifiers (indiv.-level: singling out is trivial)	*Advanced Pseudonymous Data* *Supposedly Anonymous Data*
	advanced	In addition: **Other attributes** are transformed **to protect against linkage** • Transformations include generalization, suppression, top- and bottom-coding, slicing, data swapping, and noise injection	Same as above, but: Spontaneous Recognition and linkage on other attributes is rendered more difficult or impossible (indiv.-level: singling out is trivial)	*Advanced Pseudonymous Data* *Supposedly Anonymous Data* *Successfully Anonymous Data*
aggregating identity-reduction	**basic**	**For a single disclosure, all individual-level data** is transformed such that **the resulting values relate to groups of at least C persons**	Singling out (followed by linking) possible by inference over multiple disclosures. (reconstruction attacks [↪])	*Advanced Pseudonymous Data* *Supposedly Anonymous Data* *Successfully Anonymous Data*
	disclosure control see Art. 2(4) Commission Regulation 557/2013	In addition: The aggregate values **are further protected** against known or even arbitrary singling out attacks **across multiple disclosures.**	Singling out over multiple disclosures is rendered difficult or impossible.	*Supposedly Anonymous Data* *Successfully Anonymous Data*

ULD for AnoMed Funded by Funded by the European Union NextGenerationEU

February 2024, Version 0.9.2

Feedback to research@datenschutzzentrum.de

Identity-Reduction: The Technical Perspective

3 Categories of Data

Possible Outcomes of Identity-Reduction Transformations

Disclaimer:
The data category cannot be determined from the data alone.

While there are indicators for data being personal, no technical test exists that guarantees anonymity. Data categories are therefore the result of a risk assessment which takes factors beyond just the data into account.

Data Category	Possibilities of (Re-)Identification
Fully Identified Personal Data	• **direct identification** is possible (since data is unchanged)
(Basic) ***Pseudonymous Data*** *personal data* (Recital 26 GDPR)	• direct identification is no longer possible • **only indirect identification** using **additional information** is possible
Advanced Pseudonymous Data *likely still personal data*	• direct identification is no longer possible • **even indirect identification** is rendered **difficult** or **prevented** (but with unknown success)
Supposedly Anonymous Data *likely anonymous* *but future re-identification cannot be excluded*	• **all relevant known re-identification attacks are excluded** • **thorough assessment of re-identification risk** results in low risk
Successfully Anonymous Data certainly anonymous *future practical re-identification can be excluded*	• **re-identification can be practically**[1] **excluded** • strong guarantees or thorough assessment of re-identification risk

[1] *practically* here means considering any party who can reasonably likely gain access to the data, its reasonably likely means, and taking into account technological developments.

ULD for AnoMed Funded by Funded by the European Union NextGenerationEU

Feedback to research@datenschutzzentrum.de

4 Glossary

Direct Identifier: A direct identifier is a value or value combination that is commonly known to be related to a given natural person or where a known procedure of limited effort can be used to establish such a relation. Direct Identifiers are often unique in a given context. Examples include a person's name, address, phone number, coordinates of residence, etc.

Relation to a natural person: A value is related to a natural person if, with a significant likelihood, the person has (positive relation) or has not (negative relation) a certain property described by that value.

Quasi-Identifier: A quasi-identifier is a value that is expected to be known about a natural person or easy to find out. Combinations of quasi-identifiers are often unique for a majority of persons. Examples include age, gender, and place of birth.

Singling Out: Singling out is a processing step executed on a data set that, for at least one data subject, results in some data value that is related to a (possibly unknown) person. Such processing can be a trivial lookup in the data set or require sophisticated inference that possibly uses additional information. Singling out through inference can also require the combination of multiple data sets as for example used in reconstruction attacks of statistical data[↪].

Inference: Inference is the process of deriving information from a data set that is not evident. Inference typically applies knowledge of functional dependencies between values, known correlations, known probability distributions, or other dependencies of values that can be expressed with models (including machine learning models). Types of inference include *attribute inference* where the result of the inference are new values that are related to the same data subject, and *membership inference* where, based on some known values of a person, it can be established that this person is indeed a data subject.

Linkage: Linkage is the process of establishing a relation between a singled-out value and an actual natural person. Simple forms of linkage *match* combinations of values of the data set with an external data set that contains direct identifiers. More sophisticated forms of linkage match on values derived by inference or use inference without matching. Linkage is only possible if at least one value relating to the data subject can be singled out.

Matching: Matching is a kind of Linkage based on comparison. The comparison can be based on equality of invariant values or the similarity or closeness of values that change.

Spontaneous Recognition: Spontaneous recognition is a kind of Linkage in which a human observer of a data set matches a singled out combination of values to the known values of a familiar person (relative, colleague, acquaintance, etc.). It uses additional information about the data subject that is knowledge much rather than materialized as data.

Aggregation: Aggregation is a mapping from values relating to multiple persons to a value that relates to a group of persons. Examples include statistics, machine learning models, and decision trees.

Genaralization: Generalization maps values to a coarser scale of measurement such that the number of possible values is reduced. Examples include re-classification of nominal values and the definition of intervals of ordinal, ratio or interval values. Genealization can involve multiple values as in mapping weight and height into a body mass index or mapping possible coordinates to districts or zones.

Suppression: Suppression eliminates values from the data set. This can be a single (for example exceptional) value, all values (i.e., a record) of a given data subject, or an attribute for all data subjects.

Top- and Bottom-Coding: Top- and Bottom-Coding is a transformation in which all values above or below a certain threshold are mapped to the same output value that represents (e.g., "above 220 cm")

Noise Injection: Noise injection is a transformation that adds random noise to data values.

Slicing: Slicing is a transformation that splits a high-dimensional data set into multiple lower-dimensional ones.

Data swapping: Data swapping is a transformation in which values belonging to different data subjects (typically belonging to some group) are swapped.

10.2 Pseudonymization Terminology

The following includes the developed Pseudonymization Terminology in (almost) original size. It consists of two A4 pages that can easily be printed front and back on a single handout sheet. The attached version represents the status as of February 2024. As an update and further development is desired, the terminology is published under a creative commons license. Future feedback may result in updated or consolidated versions.

ULD for AnoMed Gefördert vom Funded by the European Union NextGenerationEU

Pseudonymization Terminology for Policy Makers

Version 0.19

pseudonymization: (defined in Art. 4(5) GDPR)
A **manner of processing** in which directly identifying data elements (*additional information*) are kept separate and protected against unauthorized use in order to prevent the identification of data subjects during the processing of *pseudonymous data*.

data pseudonymization:
Data pseudonymization is a transformation of fully identified *personal data* that separates *pseudonymous data* and *identifying information*

pseudonymous data:
Pseudonymous data is *personal data* in which data subjects cannot be identified without the use of *additional information*; *identifying data* that results from *data pseudonymization* is one kind of *additional information*.

additional information: (defined in Art. 4(5) GDPR)
Additional information is any information suited to be combined (typically by linked) with *pseudonymized data* in order to identify (at least some) data subjects. One kind of *additional information* is the *identifying information* that results from *data pseudonymization*; other kinds of suitable *additional information* can exist and be held either by the controller or by external parties.

identifying information:
Identifying information is a kind of *additional information* that is the result of *data pseudonymization* and is kept separately and protected during *pseudonymization*. It permits to establish a one- or bi-directional relation between fully identifying data elements and the *pseudonyms* used in *pseudonymous data*.

pseudonymization reversal information: (used in Art. 44(3) EHDS)
Pseudonymization reversal information is bi-directional *identifying information* that permits to map from *pseudonyms* to fully identifying data elements.

pseudonymization reversal: (used in Art. 44(3) EHDS)
Pseudonymization reversal is the inverse of *data pseudonymization* that maps *pseudonymous data* plus *pseudonymization reversal information* to fully identified personal data.

reversible pseudonymization:
Reversible pseudonymization is *pseudonymization* in which the *pseudonymization reversal information* is kept available to enable a full or partial *pseudonymization reversal*.

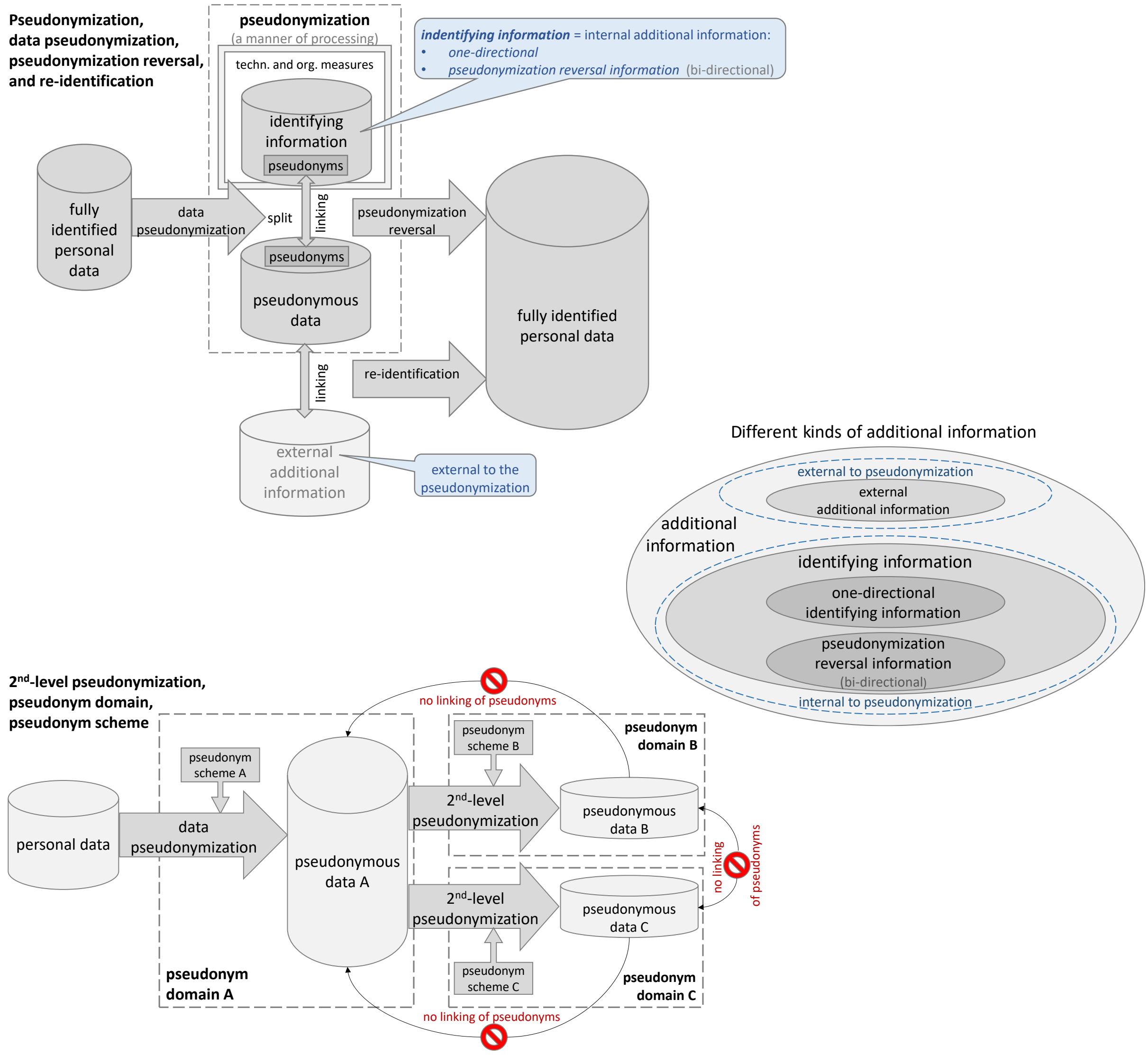
irreversible pseudonymization:
Irreversible pseudonymization is *pseudonymization* in which the *pseudonymization reversal information* is not or no longer kept such that the controller is unable to perform a full or partial *pseudonymization reversal*.

pseudonym:
A *pseudonym* is a handle for data subjects used on both the *pseudonymous data* and the *identifying information*.

pseudonym scheme:
A *pseudonym scheme* is the manner in which *pseudonyms* are created during *data pseudonymization*.

pseudonym domain: The context in which a single *pseudonym scheme* is applied and consequently, each data subject is identified by a unique pseudonym that allows linking of data elements belonging to the same data subject.

2nd-level pseudonymization:
2nd-level pseudonymization is a transformation that replaces the *pseudonyms* in *pseudonymous data* with newly created ones. It uses a separate *pseudonym scheme* to create a distinct ("unlinkable") *pseudonym domain*.


Pseudonymization,
data pseudonymization,
pseudonymization reversal,
and re-identification
pseudonymization
(a manner of processing)
techn. and org. measures
identifying information
pseudonyms
indentifying information = internal additional information:
• one-directional
• pseudonymization reversal information (bi-directional)
fully identified personal data
data pseudonymization
split
linking
pseudonyms
pseudonymous data
pseudonymization reversal
fully identified personal data
linking
re-identification
external additional information
external to the pseudonymization
Different kinds of additional information
external to pseudonymization
external additional information
additional information
identifying information
one-directional identifying information
pseudonymization reversal information
(bi-directional)
internal to pseudonymization
2nd-level pseudonymization,
pseudonym domain,
pseudonym scheme
no linking of pseudonyms
pseudonym scheme A
pseudonym scheme B
pseudonym domain B
personal data
data pseudonymization
pseudonymous data A
2nd-level pseudonymization
pseudonymous data B
no linking of pseudonyms
2nd-level pseudonymization
pseudonymous data C
pseudonym scheme C
pseudonym domain A
pseudonym domain C
no linking of pseudonyms