AnoMed

Kompetenzcluster Anonymisierung für medizinische Anwendungen

Deliverable 4.9.5

Terminologies on Pseudonymization and Anonymization

Funded by Bundesministerium für Bildung und Forschung

Funded by the European Union NextGenerationEU

https://anomed.de

UAP	4.9.5
Date	28.10.2024
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

Table of Contents

1 Overview

The present deliverable describes a second round of terminology work within work package 4.9. The first round represented in D4.9.5 produced a terminology specific to pseudonymization plus graphical handouts about types, scopes, and possible outcomes of identity-reduction transformations.

The present task aims at complementing this work by focusing on deniability for data subjects. In particular, deniability if data cannot be associated with a given data subject with sufficient certainty.

To create a better understanding of what anonymous really means, almost weekly discussions between the legally-oriented team of project partner ULD and the technically/scientifically-oriented terms at UzL-Privsec and UHH-SVS were conducted as video conferences.

In this work, the terms coming from the legal side (such as *linkage*, *singling-out*, *inference, direct and indirect identification, etc.*), were compared with and related to technical concepts (such as *predicate-singling-out*, and *reconstruction* attack).

In a **first result** of the work, the improved mutual understanding was then expressed in a **state-transition diagram** about **identity-reduction concepts**. The diagram is compatible with the handouts of D4.9.4 (in both terminology and graphical representation/color scheme) and puts emphasis on relations (i.e, transitions between) of concepts and the possible transformations between concepts. This approach was chosen to capture the actual semantics of concepts more precisely than just a glossary consisting of terms and their (textual) definition.

A **second result** is a comprehensive **taxonomy** of how controllers can **claim** that **data is** indeed **anonymous**. It presents a list of 15 possible ways that represent legally motivated claims of anonymity. The list is designed to be comprehensive, i.e., there are no other possible claims. Also, each claim is designed to be complete. For example, it is not possible to claim anonymity only based on large equivalence classes for quasi-identifiers (i.e., a large K in K-Anonymity) without providing reasons why linkage on other attributes is prevented. The claims are ordered by their strength. For example, mathematical methods with strong guarantees are stronger than asserting a lack of motivation by possible attackers. The taxonomy is unconcerned about how claims are supported by facts, though; it solely provides an order for the logical reasoning of claims. Weakly supported "stronger claims" may thus result to be inferior to solidly supported "weaker claims".

The taxonomy can represent a terminological tool for stating policies. For example, it renders it possible for policy makers to mandate a minimal strength of argumentation in order to approve data to be considered anonymous. This permits for example to ban a line of reasoning that is unlikely to hold over time (such as that there exists insufficient additional information).

The taxonomy can also be used by controllers as meta-data to make the structure of their reasoning about anonymity explicit and thus render the analysis and verification of claims easier. For this purpose, the taxonomy contains checkboxes that can be used to mark all applicable claims (which then have to be supported by additional text beyond the taxonomy).

By allowing to check the boxes of multiple claims, the taxonomy permits the use of "multiple lines of defense". For example, the main (i.e., strongest) reason for anonymity may be the aggregation of data across multiple data subjects (i.e., statistics). For the case that this protection against identification should fail (namely through a re-identification attack), the difficulty to find suitable additional information for identification by linkage represents a second line of defense.

2 Identity-Reduction State-Transition Diagram

The first major result of task 4.9.5 is an identity-reduction state-transition diagram. It is described in more detail in this section. The identity-reduction state-transition diagram attempts to use a graphical instead of textual form of communication to depict relevant concepts and their relationships.

The diagram is based on the notion of anonymity used in the GDPR. Here, anonymous is defined in Article 4(1) in terms of its contrary, namely *personal data*. Personal data directly or indirectly related to an identified or identifiable natural person. These two possibilities are visualized in the following figures.

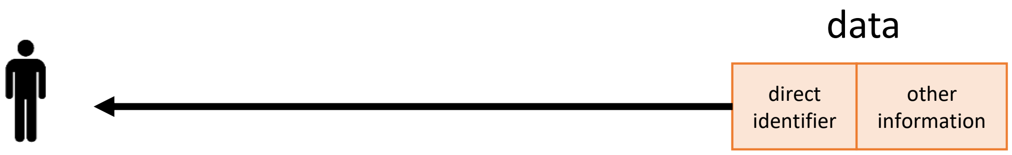


Figure 1: Direct identification of a data subject.

Art. 4(1) GDPR:
'personal data' means any **information relating to** an **identified** or identifiable natural person […] who can be identified, **directly** or indirectly, in particular by **reference to a** [direct] **identifier** […] or to one or more factors specific to […] that natural person;

Figure 1 shows the case of direct identification. Here, the related person can be directly found out based on the data themselves. Consequently, the data must contain some directly identifying elements, called *direct identifiers*. This direct identifier and other information (typically organized as a "data record") can thus be associated to the related person. The data must thus be "individual-level" data (since such association would not be possible for data where values are aggregated from values pertaining to multiple persons).

The text box below the figure states the relevant legal text of Article 4(1). The highlighting, omissions (marked by […]) and explanative additions (e.g., [direct]) were added by the author.

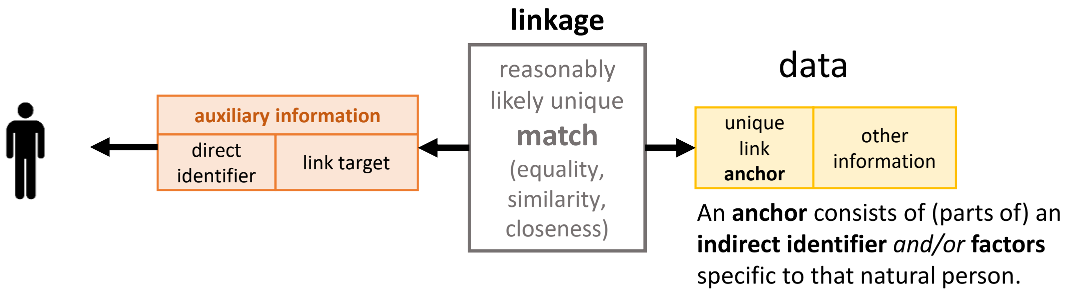

linkage
reasonably likely unique
match
(equality, similarity, closeness)
auxiliary information
direct identifier
link target
data
unique link anchor
other information
An **anchor** consists of (parts of) an **indirect identifier** *and/or* **factors** specific to that natural person.

Figure 2: Indirect identification of a data subject that requires auxiliary information.

Art. 4(1) GDPR:
'personal data' means any **information relating to** an identified or **identifiable** natural person […] who can be identified, directly or **indirectly**, in particular by **reference to an** [indirect] **identifier** […] **or** to **one or more factors** specific to […] that natural person;

Fehler! Verweisquelle konnte nicht gefunden werden. shows the case of indirect identification. Here, identification of the related person is only possible with the use of auxiliary (or additional) information[1]. This is interpreted in the following way. The data fails to contain direct identifiers. Instead it contains a data element or a combination of elements that is/are unique in the individual-level data set. This uniqueness permits to single out the data record from the data set (i.e. distinguish the data elements that pertain to a single person). We call the unique data element(s) an "anchor".

The auxiliary information is necessary to identify the person uniquely singled out by an anchor. To be suitable for this purpose, the auxiliary information must contain directly identifying data elements (i.e., direct identifiers). In addition, it must be possible to to associate a record of the auxiliary data to this anchor. This is possible to some kind of "matching" that establishes such association. The simplest kind of matching is a "join" based on the equality of values. In other words, the auxiliary data must contain the value of the anchor. The equivalence then establishes a certain association.

In more complex kinds of linking, the match could be established based on similarity or closeness. This results in more or less likely associations. To consider an association to establish identification, the likelihood must be reasonably likely.

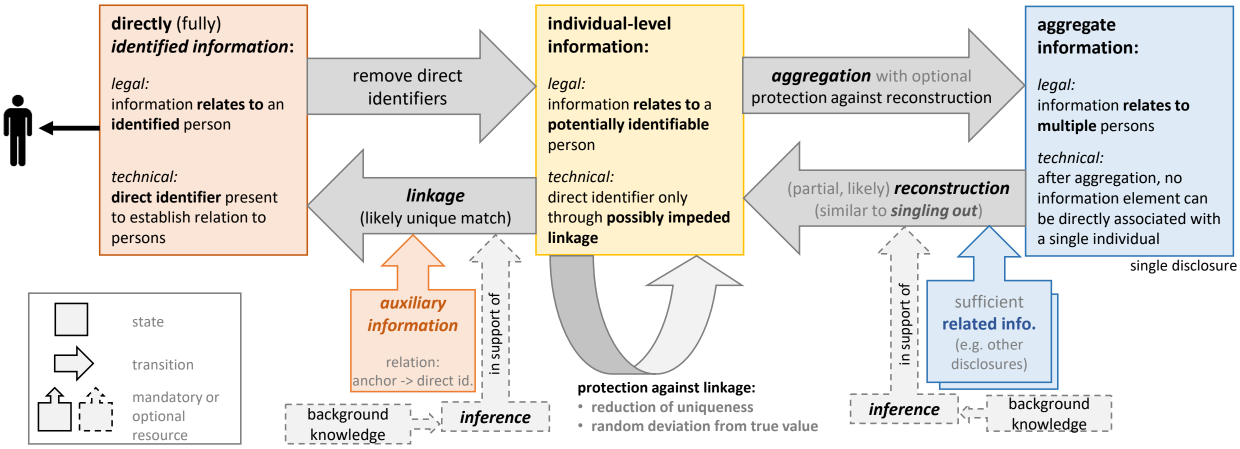

directly (fully) identified information:
legal: information relates to an identified person
technical: direct identifier present to establish relation to persons
remove direct identifiers
individual-level information:
legal: information relates to a potentially identifiable person
technical: direct identifier only through possibly impeded linkage
aggregation with optional protection against reconstruction
aggregate information:
legal: information relates to multiple persons
technical: after aggregation, no information element can be directly associated with a single individual
single disclosure
linkage
(likely unique match)
(partial, likely) reconstruction
(similar to singling out)
auxiliary information
relation: anchor -> direct id.
in support of
background knowledge
inference
protection against linkage:
• reduction of uniqueness
• random deviation from true value
in support of
sufficient related info. (e.g. other disclosures)
inference
background knowledge
state
transition
mandatory or optional resource

Figure 3: State-Transition Diagram of Identity-Reduction Concepts. (See Appendix for larger version).

This discussion of direct and indirect identification is the basis for the design of the state-transition diagram shown in Figure 3. In particular, the diagram shows that moving from directly (fully) identified information to only indirectly identifiable information is performed by removing direct identifiers. The opposite transition is only possible with the use of auxiliary information that is used to establish a *linkage*. Linkage is one of the three terms use by the Article 29 Data Protection Working Party to assess whether data is indeed anonymous (i.e., identification of data subjects is impossible)[2].

[1] The interpretation, that indirect identification requires auxiliary (additional) information is also motivated by the GDPR's definition of pseudonymization in Article 4(5).

[2] ARTICLE 29 DATA PROTECTION WORKING PARTY, WP216, Opinion 05/2014 on Anonymisation Techniques, Adopted on 10 April 2014, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (last visited 14/10/24).

Linkage can be supported by inference. For example, in order to produce a match, functionally derived values may be derived in order to produce the linkage. Similarly, strong correlations could be used to identify likely matches. Also, while no record of the data can be uniquely matched to additional information, knowing that one value is the same for a whole group, may render linkage of this single value possible[3].

In addition to the transitions between directly identified information and individual-level information (discussed above), there is a wide range of transformations of individual-level information to individual-level information with reduced potential of identification. In particular, these identity-reduction transformations aim at reducing the potential for successful linkage.

Linkage protection can be achieved in two ways:

- Reducing the uniqueness of value combinations in the data (mostly records), and
- Introducing (random) deviations from the true values such that matches to auxiliary information become less certain or fail.

A prime example of the former type of reducing uniqueness is K-Anonymity which uses suppression and generalization to avoid uniqueness in the combinations of parts of the attributes (the so called "quasi-identifiers). The uniqueness of data records of any kind of attribute combinations can be measured with the "Special Uniques Detection Algorithm" (SUDA)[4] which is known from statistical disclosure control. Detecting special uniques can be used as a factor to assess the linkage potential of data (records).

The latter type of introducing random deviations is less common in practice. It is implemented by epsilon-differential privacy methods applied to individual-level data sets. Its relevance in practice is very low though, since usually, so much noise needs to be injected to reach the desired privacy goal that the data utility is unfit for practical purposes.

Transformations that protect against linkage typically just lower the probability of successful linkage. They usually cannot guarantee that linkage is practically impossible. Using these transformations to support the claim of anonymity of a data set is only possible with additional assumptions about either the available auxiliary information or the capability of possible attackers.

The third kind of transitions maps between the states of individual-level and aggregate information.

In one direction, this includes transformations that *aggregate* individual-level value to aggregated values that pertain to multiple individuals. Prime examples for such aggregations are statistics, AI models, and synthetic data. If the aggregation is effective, no single data value can be attributed to a single person.

In the other (i.e., inverse) direction, the transformations are called *reconstructions*. They typically require a multitude of aggregated data sets (i.e., disclosures) to build equation systems that can be solved for individual-level input values of the aggregation. A large-scale reconstruction has been

[3] In the context of K-Anonymity, this I known as a "homogeneity attack".

[4] See Elliot, M. J., Manning, A. M., & Ford, R. W. (2002). A Computational Algorithm for Handling the Special Uniques Problem. International Journal of Uncertainty, Fuzziness and Knowledge Based System , 10 (5), 493-509, and https://pypi.org/project/suda/ for an open source implementation and illustrative examples (last visited 14/10/24).

demonstrated[5] by the U.S. Census Bureau base on their 2010 statistical census data that previously underwent rigorous statistical disclosure control.

Were the aggregation is not an effective protection against singling out, at least for some data subjects, it may also be possible to reconstruct individual-level attributes from a single data set. This is called *model inversion*[6]. Due to its ineffective protection against (at least partial) singling out, it can be considered to be a kind of *obfuscation* of individual-level data much rather than an effective aggregation (in analogy to encryption). The state-transition diagram therefore only considers effective aggregation.

Obfuscated individual-level data and effectively aggregated data may co-exist in the same data set where individual-level information is only obfuscated for a few (exceptional) data subjects (e.g. outliers) while effectively aggregated for other (more average) data subjects. Aggregation is typically more effective where multiple similar data subjects "cluster" together and less (or not) effective for individuals who are dissimilar to other individuals.

An effective aggregation protects against singling out of information elements that pertain to an individual person. *Singling out* is mentioned in the GDPR's Recital 26 as a possible means for identifying data subjects. Singling out is one of the three terms use by the Article 29 Data Protection Working Party to assess whether data is indeed anonymous (i.e., identification of data subjects is impossible)[7]. While singling-out is impossible in effectively aggregated data, reconstruction of aggregated data sets converts them into a form where (at least partial) singling out is again possible. The concepts of *singling out* and *reconstruction* are therefore closely related.

Inference may support reconstruction of individual-level information. It can employ background knowledge. For example, it may use knowledge about functional dependencies or strong correlation to extract information from certain disclosures to render them suitable to contribute to the solution of a reconstruction equation system.

In summary, the state-transition diagram captures the semantics of the three concepts *linkage*, *singling out*, and *inference* that were used[8] by the Article 29 Data Protection Working Party to describe anonymity. It shows how *linkage* and *singling out* are the prime concepts that transition between two major states of information, and that *inference* has only a supportive role for these prime concepts. In other words, inference that fails to support the two prime concepts is unable to contribute to identification.

The state-transition diagram also shows that *linkage* is always necessary for identification. The ability of *singling out* is a prerequisite for linkage and thus identification. Aggregated information that prevents singling out without the use of additional disclosures introduces an additional protection against identification. In particular, identification then requires two steps: reconstruction and linkage.

[5] Garfinkel, Simson & Abowd, John & Martindale, Christian. (2019). Understanding database reconstruction attacks on public data. Communications of the ACM. 62. 46-53. 10.1145/3287287. https://dl.acm.org/doi/pdf/10.1145/3291276.3295691 (last visited 16/10/24).

[6] Veale M, Binns R, Edwards L., 2018 Algorithms that remember: model inversion attacks and data protection, law. Phil. Trans. R. Soc. A 376: 20180083. https://doi.org/10.1098/rsta.2018.0083 (last visited 16/10/24).

[7] See footnote 2.

[8] See footnote 2.

3 Taxonomy of Anonymity Claims

The second major result of task 4.9.5 is a taxonomy of possible anonymity claims. It is described in more detail in this section. Its development was based on the state-transition diagram described above.

3.1 Problem addressed by the taxonomy

In many contexts, it is necessary **to assess the risk of re-identification** of identity-reduced data. If this assessment results in an **insignificant ris**k, the data is **(legally) considered to be "anonymous"**. This fact underlines the importance of such risk assessment.

Currently, no generally accepted or even standardized approach to such risk assessment exists. Consequently, a wide range of approaches exist in practice. It is often not clear how to compare risks assessed with different approaches. It may also be the case that a given approach focusses on a certain aspect and neglects others[9].

The assessment of re-identification risk of data, mostly resulting in the decision of whether data is indeed "anonymous", is crucial for a range of actors including the following:

- Primarily, **actors who process data** need to decide whether this data is indeed anonymous and thus falls outside the GDPR or whether the risk of re-identification is significant enough to apply the requirement that the GDPR imposes on controllers.
- **Supervisory Authorities** and **courts** who oversee compliance with the GDPR have to be able to make and/or verify such assessments.
- **Policy makers** need to express how much risk they tolerate in a given setting. For example, in the EHDS, Health Data Access Bodies must issue policies on when a result of data analysis is considered to be legally anonymous and can be published on their website[10].

Currently, assessment results, laws and policies can only use a very restricted vocabulary: "anonymous" or "not anonymous" (in a legal sense). Expressing legal requirements and policies solely with the single word "anonymous" leads to certain **difficulties** that **include** the following:

- The **interpretation of** the single term of **"anonymous" is left to each actor** who assesses the risk of re-identification. Consequently, a wide variety of interpretations exists and not all interpretations can be considered to be correct.
- The acceptable risk can be expressed solely by a single state represented by the term "anonymous". Since risk is typically understood as a continuum and acceptable risk depends on the circumstances, a **more diversified and precise vocabulary** for expressing acceptable risk is desirable.

[9] For example, an given risk assessment may focus on a large K achieved in K-anonymous data, but neglects to assess the possibly substantial identification risk posed by unique combinations of attributes which are not quasi-identifiers.

[10] See Article 46(11) of the Commission Proposal of the EHDS. Also note that health data access bodies are considered to be joint controllers together with data users (see Art. 51 EHDS Proposal).

To address these difficulties, a taxonomy on how actors can reason that data is indeed anonymous is proposed. The main objectives and non-objectives of this taxonomy are described in the following.

3.2 Objectives and Non-Objectives of the Taxonomy

The taxonomy distinguishes 15 distinct ways in which one can logically reason that data is anonymous. The arguments are arranged in order of their strength of argument. This order does not necessarily imply that data supposed to be anonymous based on a stronger argument necessarily possesses a lower risk of re-identification; the actual risk depends on the actual elements used in the logical chain of reasoning.

The taxonomy attempts to satisfy the following objectives:

1. The context and target of the claim must be explicit and clear.
2. Each argument for anonymity must be **complete** by including the entirety of reasoning elements to conclude that the data is indeed anonymous.
3. The taxonomy attempts to be **comprehensive** in a way that it enumerates all possible lines of reasoning that data is anonymous.
4. The scope of the taxonomy is limited to **the logical structure of argumentation**; the strength of individual elements of argumentation and of the overall argument are out of scope.
5. The taxonomy supports **multiple lines of defense** in as far as it allows to identify a main argument for the data to be anonymous and indicate additional, weaker, arguments for successful anonymity that apply even in the case where all the stronger arguments fail.

The taxonomy does **not attempt** to fulfil the following objectives:

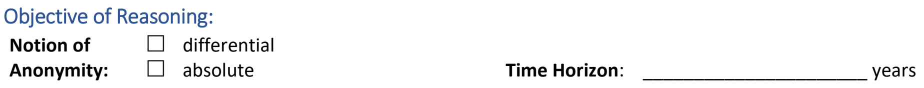
- While the taxonomy guarantees that the structure of reasoning is valid, it provides no such guarantees for the actual reasoning. Examples for invalid reasoning include guaranteed properties of insufficient strength (such as a k of 2 in k-anonymity or an epsilon of 100 in differential privacy), or erroneous reasoning (such as that suitable auxiliary data fail to exist or no possible attacker is sufficiently motivated).
- While the taxonomy provides a richer vocabulary to express what re-identification risk is adequate in a given situation, it does not attempt to answer the question of which lines of reasoning are acceptable in which situation or to satisfy the legal notion of anonymity.
- While the taxonomy attempts to provide a comprehensive set of lines of reasoning, it does not imply that every line is actually applicable in practice. It may be that certain lines of reasoning (such as 1.3) apply only in rare cases where likely the utility of the corresponding data is below the threshold of practical use.

3.3 Approach to fulfil the Objectives

The following describes how the above objectives have been reached.

3.3.1 Well-Defined Context and Target of Claims

Anonymity claims are made for an assumed context and pursue a specific target. To make this explicit, the handout includes the section shown in Figure 4.



Objective of Reasoning:

Notion of Anonymity: □ differential □ absolute

Time Horizon: ____________________ years

Figure 4: Defining the Context of the Claims.

This section lets users declare the pursued target, i.e., notion (or definition) of anonymity. Namely, this is a choice between a differential and an absolute definition. The detailed explanation of the two notions is provided in the appendix under 5.1.

The second element that is of importance to specify the context of the claims is the assumed time horizon. To make this explicit is important since several aspects of a claim can change over time. Prime examples for this is available auxiliary information for linkage, known attacks for certain kinds of protections, and computing capabilities for example to compute large-scale equation systems for reconstruction.

3.3.2 Completeness

Completeness means that the complete set of logical elements to reason that data is anonymous are present. In the table, such a complete set of logical elements is represented by a horizontal traversal from left to right. The traversal then touches all necessary logical elements.

In a more populistic analogy, completeness of arguments prevents the possibility to brag about the titanium lock on the front door while leaving all windows open; or planting a reinforced steel pole in the middle of the street and hope that the car hits there.

The most common identity-reduction transformations provide only partial guarantees of a certain strength that identification is not possible. Such guarantees have to be complemented with additional reasoning in order to be complete. For example, epsilon-Differential-Privacy provides guarantees of strength epsilon but has to be complemented with a sound management of the privacy budget. Similarly, k-anonymity provides a guarantee of strength k that linkage to quasi-identifiers is not possible; for completeness this has to be complemented with reasons why linkage on non-quasi-identifier attributes[11] is not possible.

In all cases, also inference in support of linkage or reconstruction needs to be considered. Such consideration can be limited to certain cases of the taxonomy, however. In particular, inference in support of reconstruction must only be considered where the reasoning is based on the impossibility of reconstruction (namely 2.1.3); similarly, inference in support of linkage must only be considered where the reasoning is based on the impossibility of linkage (namely 1.2).

[11] More precisely, this must include also combinations of quasi-identifiers and non-quasi-identifier attributes that can be used as anchors for linkage.

By complementing the partial guarantees of common identity-reduction transformations and considering inference where necessary, the taxonomy of reasoning attempts to be complete.

3.3.3 Comprehensiveness

The requirement of comprehensiveness means that all possible cases are covered by the taxonomy of reasons.

The approach to reaching logical comprehensiveness is partly based on the underlying state-transition diagram for identity-reduction and on using a set of cases and its complement. In the latter case, the union of a set and its complement always covers all possible cases.

The comprehensiveness of the overall structure is based on the state-transition diagram of identity reduction. Here, data is legally anonymous if identification is not possible. Identification can only be prevented in two ways: (i) by preventing reconstruction or (ii) by preventing linkage. By considering both possible reasons for anonymity, the comprehensiveness is reached.

Within the reconstruction argument, the set of cases "with mathematical guarantees" (2.2) and its complement "without mathematical guarantees" (2.1) are considered and comprehensiveness is thus reached. Also the two options 2.2.1 and 2.2.2 represent a set of cases and its complement. The same goes for 2.2.1.1. and 2.2.1.2. In the latter case, significant disclosures do not exist, in the former case (i.e., its complement), they do exist and for the data to be anonymous, it must be impossible that they are used for identification.

The subcases of 2.1 follow a similar scheme of complementary sets: in 2.1.3, reconstruction is not possible. In 2.1.1 and 2.1.2, it is possible (i.e., the complement set) and identification is prevented using other arguments. In 2.1.2, additional disclosures are not available to be used for identification; in 2.1.1 they are available, but are not used based on reasoning about potential attackers.

As illustrated, the table therefore satisfies the comprehensiveness requirement for the reconstruction part.

The initial comprehensiveness of the linkage part is based on the model of indirect identification where linkage consists of establishing a (reasonably likely, unique) relation between (i) a link target contained in auxiliary information and (ii) a link anchor contained in the data that are the subject of the reasoning about anonymity. For comprehensiveness reasons, such a relation cannot be established either since (i) no suitable link targets are used for identification or (ii) no suitable link achors are available to establish matches to targets. This is reflected in case 1.3 where no matter the possibly existing auxiliary information (and its targets), there are no suitable anchors. Cases 1.1 and 1.2 then assumes that the data set contains suitable anchors, but the relation cannot be established for other reasons: For 1.2, this reason is that it is impossible to obtain suitable auxiliary information to identify the data; for 1.2, suitable auxiliary information is available, but it is assumed that potential attackers were unable or unmotivated to use it for identification.

The above reasoning shows that also the linkage part of the table satisfies the completeness requirement.

3.3.4 Multiple Lines of Defense

The taxonomy of reasoning about anonymity provides the structure of reasoning, not the actual arguments. In other words, it presents the necessary conclusions of a reasoning element, not the arguments that are used to reach those conclusions. The taxonomy thus cannot provide any guarantee, that the reasoning is valid. In particular, arguments may be “optimistic” by making assumptions that fail to apply in reality or they may even be logically flawed.

It is likely impossible to reason about anonymity without any assumptions about other players (who might unknowingly erode the privacy budget), the state of the art of re-identification, available auxiliary information, or the capabilities and motivations of potential attackers (the attacker model).

Consequently, it is always possible that the main reason for successful anonymity is later discovered to be flawed. In this case, for the argumentation it is relevant to have multiple lines of defense. For example, assume that the main (and strongest) reason for anonymity is the impossibility of reconstruction of aggregate information. Then, additional lines of defense are represented by reasons why linkage is not possible, even in the case where the main protection unexpectedly failed.

In order to support multiple lines of defense against identification, the arguments in support of anonymity (i.e., lines in the table) are ordered by their logical strength.   Higher up in the table and a higher number are then stronger arguments. The table allows to mark multiple arguments in support of anonymity by checking multiple boxes in the right-most column. Due to the ordering, the top one is then considered the strongest and the additional ones are additional lines of defense against identification.

4 Conclusions

The present deliverable has shown how the frequent exchange between the involved project partners has resulted in two major terminology products: a state-transition diagram of identity-reduction and a taxonomy for anonymity claims.

The graphical presentation and the concise handout format targets busy policy makers.

The reported versions are intended to evolve further over time. An internal review and approval process at ULD has been started. Further dissemination outside of the project is planned.

5 Appendix

The Appendix contains an essay on the two notions of anonymity, as well as the two developed terminologies in full size.

5.1 Two Notions of Anonymity

5.1.1 Relevance

In the context of the European Data Strategy and the emerging data spaces, a clash of concepts and resulting potential conflict is foreseeable. In particular, the concepts of data science and data protection could collide since they see the same data as an important analysis result (a pattern in the data) or as an attack on privacy (e.g., a homogeneity attack), respectively.

In an attempt to avoid a possible conflict, this essay (i) raises awareness that there are two possible notions of "anonymous" on the data protection side, and (ii) that the compatibility with the concepts of data science depend on the choice of notion. It encourages a discussion that could constitute an important input to the EDPB who is currently writing guidelines on anonymization[12].

5.1.2 Two definitions of "anonymous"

In the literature, there are two notions of anonymity. Their definition and sources are provided in this section. The fact that there are two distinct notions of anonymity is for example stated by D. Smith[13].

The first notion is the most commonly used one. While many sources for a definition could be used, the following uses one from the area of statistical disclosure control as it is used by the official statistics bodies of Member States and the European Union. The definition is not that of *anonymity*, but of its contrary, i.e., *disclosure*.

ESSNet S D C[14], A Network of Excellence in the **E**uropean **S**tatistical **S**ystem in the field of **S**tatistical **D**isclosure **C**ontrol, defines Disclosure as follows[15]: "*A disclosure occurs when a person or organisation recognises or learns something that they did not know already about another person or organisation, via released data.*"

Based on this definition of *disclosure*, the first notion of *anonymous* is defined as follows:

> Definition: ***Anonymous (1)***
> *A data set is anonymous if it is not possible to learn anything new about a data subject from it.*

[12] 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.

[13] See Section 2.3.2, 2nd paragraph in Smith, Duncan. (2019). Re-identification in the Absence of Common Variables for Matching. International Statistical Review. 88. 10.1111/insr.12353. https://doi.org/10.1111/insr.12353.

[14] https://cros-legacy.ec.europa.eu/page/essnet_en, last visited 18/12/23.

[15] Handbook on Statistical Disclosure Control | CROS (europa.eu)

A second definition originates in *differential privacy*[16]. In particular, Dwork et al state[17]: "*Our ultimate privacy goal when releasing information about a sensitive dataset is to ensure that anything that can be learned about an individual from the released information, can be learnt without that individual's data being included. This goal does not ensure that nothing about an individual can be learnt from the released information, which can only be achieved by releasing no information*."

On this basis, the second notion of anonymous is defined as follows:

Definition: ***Anonymous (2)***
A data set is anonymous if it is not possible to learn anything new about a given data subject from the data set that could not be learned if that data subject was not included in the data set.

5.1.3 An example data set

The following gives an example of a data set that demonstrates the difference of the notions.

Assume that in support of public health, a large data set across the population was collected about a number of diseases. It has been k-anonymized[18] and contains the following attributes for every data subject: Place of residence, gender, age, and a Boolean indicator for every disease, showing whether the data subject is affected by it. Place of residence, gender, and age are considered to be quasi-identifiers in the k-anonymization such that for each possible combination of these attributes, at least k data subjects share the same values. K has been chosen to be sufficiently large.

Further assume that the data set covers a remote mining town where all male inhabitants are employed by a mining company which unknowingly has been affected by a serious pollution of their drinking water. Consequently, the data set shows that all males of working age are affected by severe kidney problems.

5.1.4 Assessment of the example data set under the two notions of anonymous.

In this section, the above example data set is assessed whether it is anonymous under the two notions.

(i) Using the first notion, the question must be posed whether it is possible to learn something new from the data set about data subjects. Clearly, it is possible to infer from the data set, that a male person of working age who lives in the mining town suffers from kidney problems. Under the first notion of *anonymous*, the data set is therefore judged to **not** be **anonymous**. In more detail, a so-called *homogeneity attack* can be used since the k-anonymized data set fails to be l-*diverse*[19].

(ii) Using the second notion, the question must be posed whether the information inferred from the data set is different in the case that a given data subject was not represented in the data

[16] Dwork, C. (2008). An Ad Omnia Approach to Defining and Achieving Private Data Analysis. In: Bonchi, F., Ferrari, E., Malin, B., Saygin, Y. (eds) Privacy, Security, and Trust in KDD. PInKDD 2007. Lecture Notes in Computer Science, vol 4890. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-78478-4_1

[17] See Section 4.1 in Dwork, Cynthia & Smith, Adam & Steinke, Thomas & Ullman, Jonathan. (2017). Exposed! A Survey of Attacks on Private Data. Annual Review of Statistics and Its Application. 4. 10.1146/annurev-statistics-060116-054123. https://privacytools.seas.harvard.edu/files/privacytools/files/pdf_02.pdf

[18] *Sweeney, Latanya (2002). "k-anonymity: a model for protecting privacy" (PDF). International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems.* ***10*** *(5): 557–570. doi:10.1142/S0218488502001648. S2CID 361794*

[19] Machanavajjhala, Ashwin & Gehrke, Johannes & Kifer, Daniel & Venkitasubramaniam, Muthuramakrishnan. (2006). l-Diversity: Privacy Beyond k-Anonymity. ACM Transactions on Knowledge Discovery From Data - TKDD. 1. 24. 10.1145/1217299.1217300. http://www.cs.cornell.edu/people/dkifer/ldiversityTKDDdraft.pdf.

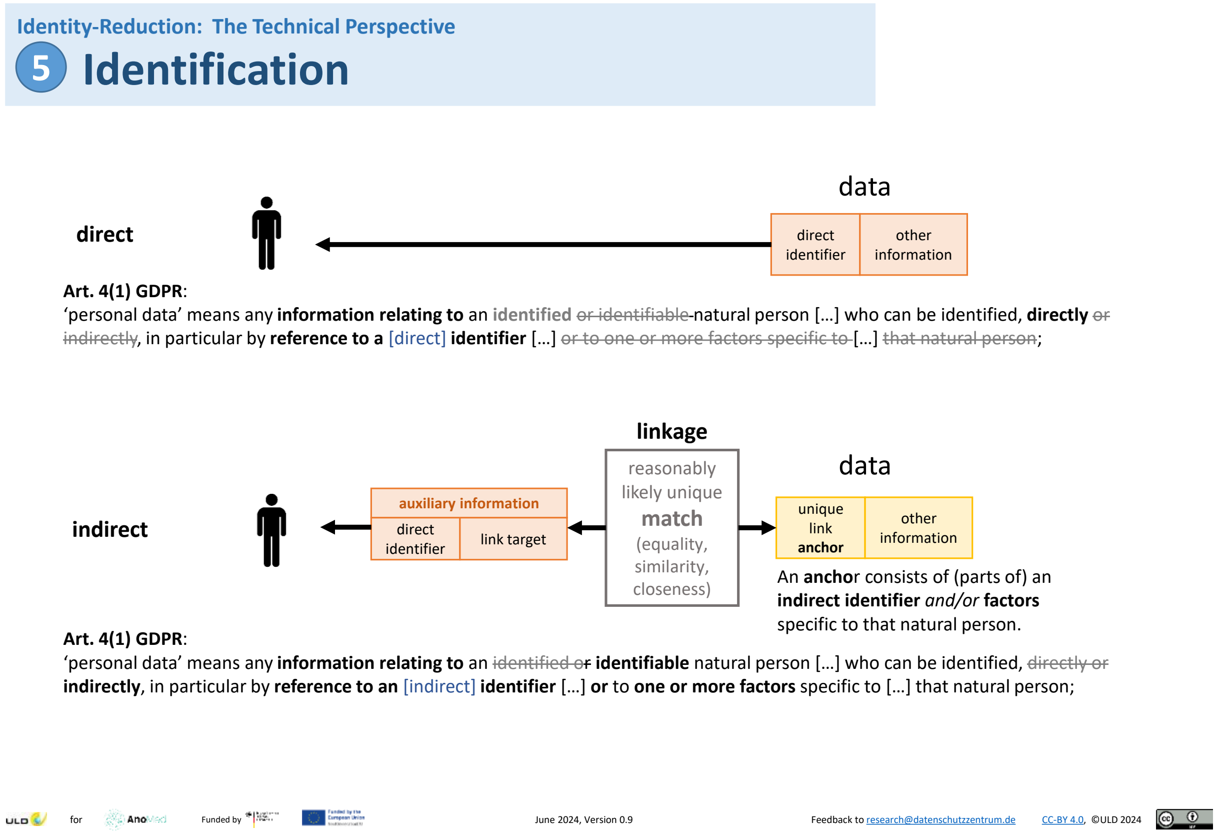
set. Clearly, the fact that working age males from that town are affected by kidney problems does not derive from a single data subject. Much rather, if one data subject is deleted from the data set, there are still k-1 data subjects left that lead to the same conclusion. Consequently, under the second notion of anonymous, the example set is indeed anonymous.

5.1.5 Need for timely clarification

A clarification of the notion of "anonymous" in data protection seems to be urgent considering the advanced plans to implement a multitude of data spaces in Europe. A timely discussion and clarification (e.g. by the EDPB in its upcoming *Guidelines on Anonymization*) could potentially avoid a situation of conflict that potentially endangers the trust of citizens to provide their personal data for the creation of data spaces.

5.2 State-Transition-Diagram of Identity-Reduction

The front page of the handout looks as follows:



Identity-Reduction: The Technical Perspective

5 Identification

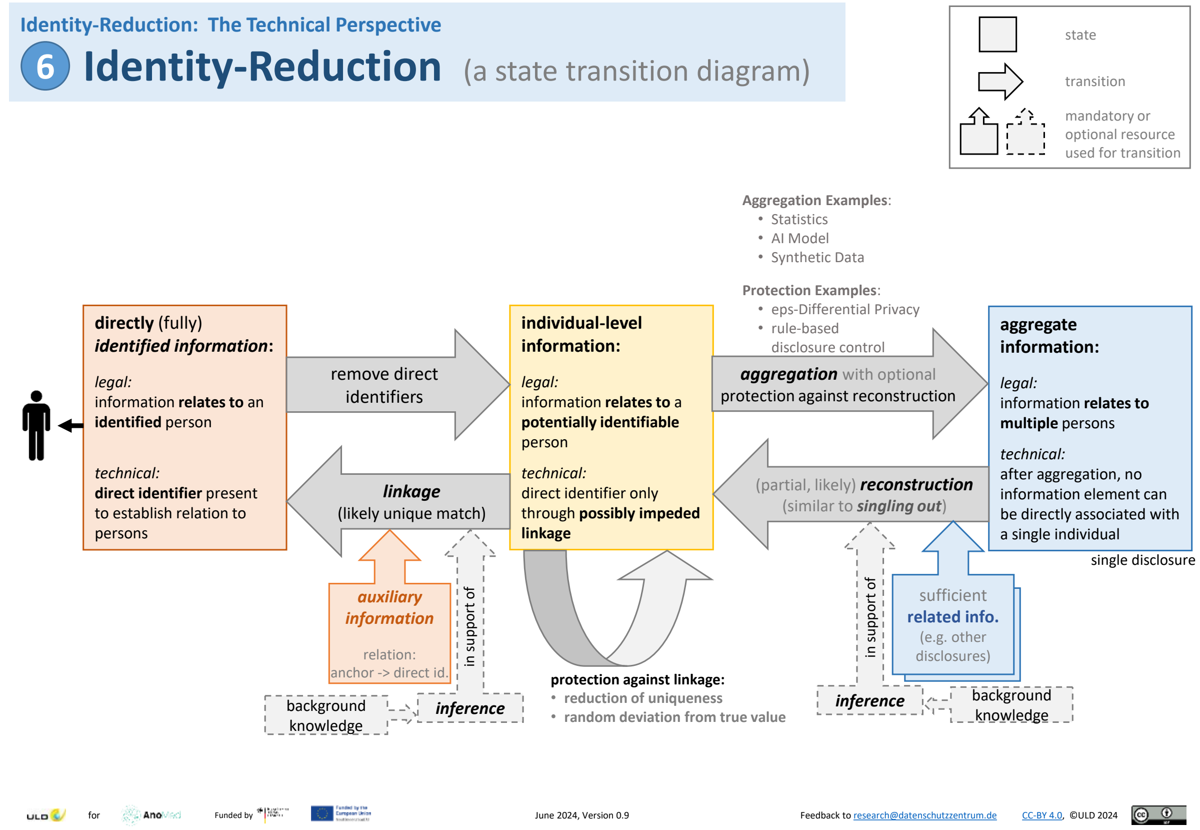
Art. 4(1) GDPR:
'personal data' means any **information relating to** an identified ~~or identifiable~~ natural person […] who can be identified, **directly** ~~or indirectly~~, in particular by **reference to a** [direct] **identifier** […] ~~or to one or more factors specific to~~ […] ~~that natural person~~;

An **anchor** consists of (parts of) an **indirect identifier** *and/or* **factors** specific to that natural person.

Art. 4(1) GDPR:
'personal data' means any **information relating to** an ~~identified or~~ **identifiable** natural person […] who can be identified, ~~directly or~~ **indirectly**, in particular by **reference to an** [indirect] **identifier** […] **or** to **one or more factors** specific to […] that natural person;

ULD for AnoMed Funded by June 2024, Version 0.9 Feedback to research@datenschutzzentrum.de CC-BY 4.0, ©ULD 2024

The following shows the back page of the handout.



Identity-Reduction: The Technical Perspective

6 Identity-Reduction (a state transition diagram)

Aggregation Examples:
- Statistics
- AI Model
- Synthetic Data

Protection Examples:
- eps-Differential Privacy
- rule-based disclosure control

directly (fully) ***identified information***:

legal: information **relates** to an **identified** person

technical: **direct identifier** present to establish relation to persons

individual-level information:

legal: information **relates to** a **potentially identifiable** person

technical: direct identifier only through **possibly impeded linkage**

aggregate information:

legal: information **relates to multiple** persons

technical: after aggregation, no information element can be directly associated with a single individual

remove direct identifiers
aggregation with optional protection against reconstruction
linkage (likely unique match)
(partial, likely) ***reconstruction*** (similar to ***singling out***)
auxiliary information
relation: anchor -> direct id.
in support of
background knowledge
inference
single disclosure
sufficient **related info.** (e.g. other disclosures)
in support of
inference
background knowledge

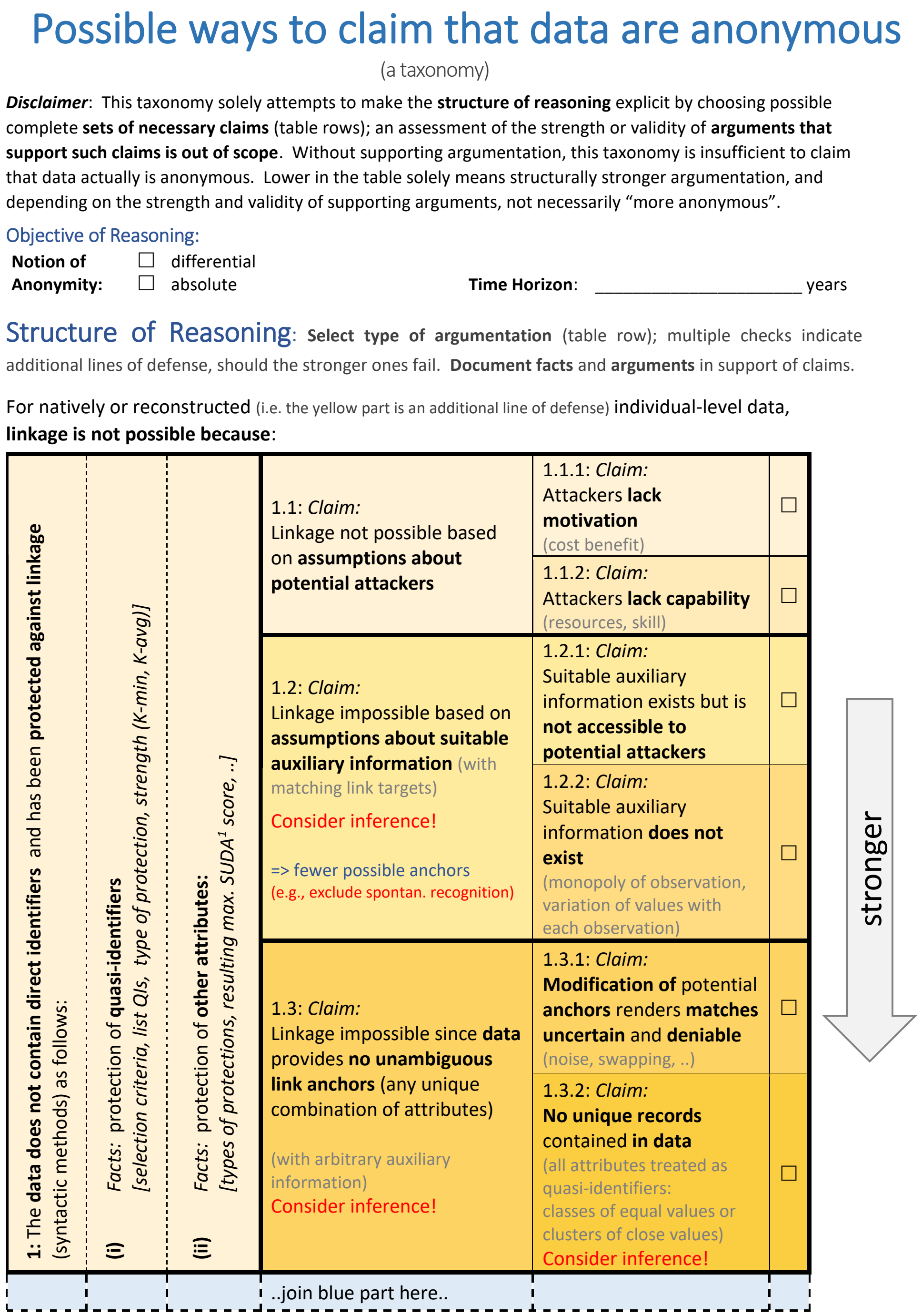
protection against linkage:
- reduction of uniqueness
- random deviation from true value

ULD for AnoMed Funded by June 2024, Version 0.9 Feedback to research@datenschutzzentrum.de CC-BY 4.0, ©ULD 2024

5.3 Taxonomy of Anonymity Claims

The taxonomy logically contains a single table. For practical purposes, it is split up in two parts that can be printed on a single double-sided sheet.

The front page looks as follows:



Possible ways to claim that data are anonymous

(a taxonomy)

Disclaimer: This taxonomy solely attempts to make the **structure of reasoning** explicit by choosing possible complete **sets of necessary claims** (table rows); an assessment of the strength or validity of **arguments that support such claims is out of scope**. Without supporting argumentation, this taxonomy is insufficient to claim that data actually is anonymous. Lower in the table solely means structurally stronger argumentation, and depending on the strength and validity of supporting arguments, not necessarily "more anonymous".

Objective of Reasoning:

Notion of Anonymity: □ differential □ absolute

Time Horizon: ______________________ years

Structure of Reasoning: **Select type of argumentation** (table row); multiple checks indicate additional lines of defense, should the stronger ones fail. **Document facts** and **arguments** in support of claims.

For natively or reconstructed (i.e. the yellow part is an additional line of defense) individual-level data, **linkage is not possible because**:

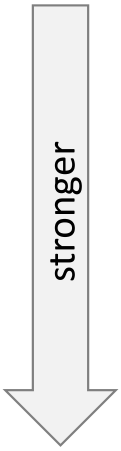
1: The data does not contain direct identifiers and has been **protected against linkage** (syntactic methods) as follows:	**(i)** *Facts*: protection of **quasi-identifiers** *[selection criteria, list QIs, type of protection, strength (K-min, K-avg)]*	**(ii)** *Facts*: protection of **other attributes**: *[types of protections, resulting max. SUDA[1] score, ..]*	1.1: *Claim:* Linkage not possible based on **assumptions about potential attackers**	1.1.1: *Claim:* Attackers **lack motivation** (cost benefit)	□
				1.1.2: *Claim:* Attackers **lack capability** (resources, skill)	□
			1.2: *Claim:* Linkage impossible based on **assumptions about suitable auxiliary information** (with matching link targets) Consider inference! => fewer possible anchors (e.g., exclude spontan. recognition)	1.2.1: *Claim:* Suitable auxiliary information exists but is **not accessible to potential attackers**	□
				1.2.2: *Claim:* Suitable auxiliary information **does not exist** (monopoly of observation, variation of values with each observation)	□
			1.3: *Claim:* Linkage impossible since **data** provides **no unambiguous link anchors** (any unique combination of attributes) (with arbitrary auxiliary information) Consider inference!	1.3.1: *Claim:* **Modification of** potential **anchors renders matches uncertain** and **deniable** (noise, swapping, ..)	□
				1.3.2: *Claim:* **No unique records** contained **in data** (all attributes treated as quasi-identifiers: classes of equal values or clusters of close values) Consider inference!	□
			..join blue part here..		

stronger

The following shows the back page:

For aggregate information, **reconstruction is not possible because:**

		..join yellow part here..		
2: The data is aggregated, and thus direct or indirect identification are only possible after successful (possibly partial) reconstruction *[type: statistics, AI-model, synthetic data, … ,minimal cell size or similar measure of aggregation level]*	**2.1:** *Fact:* **Data without mathematically guaranteed reconstruction protection** (e.g., • statistics without additional protection, • (empirical) rule-based disclosure control) *Reconstruction protection:* facts*: [type, properties]* *E.g. (none, …)*	2.1.1: *Claim*: Reconstruction is assumed to be impossible based on **assumptions about potential attackers**:	2.1.1.1: *Claim:* Attackers **lack motivation** (cost/benefit)	□
			2.1.1.2: *Claim:* Attackers **lack capability** (skill, resources, ..)	□
		2.1.2: *Claim*: Reconstruction is assumed to be impossible based on **assumptions about additional disclosures** and ..	2.1.2.1: *Claim*: **..addl. disclosures exist but are not accessible** by potential attackers	□
			2.1.2.2: *Claim*: ..Significant additional disclosures **don't exist**	□
		2.1.3: *Claim*: Reconstruction is assumed to be impossible based on current **state of the art** and .. **consider:** reconstruction protection, inferences	2.1.3.1: *Claim*: Known attacks *[enum]* fail **based on assumptions** about state of the art	□
			2.1.3.2: *Claim*: Known attacks *[enum]* fail as **verified with own data**	□
	2.2: *Fact:* **Data with mathematically guaranteed reconstruction protection** **Guarantee:** facts*: [type, strength]* *(e.g. eps-DP, eps)*	2.2.1: *Fact*: The privacy budget is managed **for own disclosures,** and..	2.2.1.1: Claim: ..no significant number of external disclosures are accessible to attackers	□
			2.2.1.2: *Claim*: ..no significant number of external disclosures exists	□
		2.2.2: *Claim*: The privacy budget is managed for both, **own and external disclosures**		□



version 0.9

---

[1] Elliot, M. J., Manning, A. M., & Ford, R. W. (2002). A Computational Algorithm for Handling the Special Uniques Problem. International Journal of Uncertainty, Fuzziness and Knowledge Based System , 10 (5), 493-509. and also https://pypi.org/project/suda/.

ULD for AnoMed Funded by