AnoMed
Kompetenzcluster Anonymisierung für medizinische Anwendungen

Deliverable 4.9.7

Strategies to Render Health Data Usable

SPONSORED BY THE
Federal Ministry
of Education
and Research

Funded by
the European Union
NextGenerationEU

https://anomed.de

UAP	4.9.7
Date	30.01.2025
Version	1.0
Status	Final
Distribution	PU
Lead Contributors (© by affiliation)	Bud P. Bruegger (ULD)
Additional Contributors (© by affiliation)	Esfandiar Mohamadi (UzL-Privsec) (Section 5.4.5.1)
Reviewers	Harald Zwingelberg (ULD)
License	CC-BY 4.0

Disclaimer:

The current deliverable attempts to contribute to the understanding of legal requirements and possible technical and organizational measures for secondary uses in data spaces.

It is hoped that the presented analysis may be helpful to controllers, authorities implementing data spaces, and supervisory authorities when guiding such entities.

In no way does this document attempt to relieve or partially relieve controllers of their responsibility to determine the purposes and means of their specific processing activities nor to assess the associated risks or determine adequate mitigation measures.

Neither ULD nor the authors have any interest in any particular processing activity nor the deployment of any particular measure.

English: Funded by the European Union – NextGenerationEU. The views and opinions expressed are solely those of the author(s) and do not necessarily reflect the views of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

German: Finanziert durch die Europäische Union – NextGenerationEU. Die geäußerten Ansichten und Meinungen sind ausschließlich die des Autors/der Autoren und spiegeln nicht unbedingt die Ansichten der Europäischen Union oder der Europäischen Kommission wider. Weder die Europäische Union noch die Europäische Kommission können für sie verantwortlich gemacht werden.

https://next-generation-eu.europa.eu/index_de

Table of Contents

1 Introduction

Data are often considered "the new oil" of society and in Europe; there is a strong desire to harvest the significant potential of data for the benefit of society. This desire has been rendered concrete in the *European Strategy for Data*[1]. This strategy has and will be implemented by a series of legal acts (see Deliverable D4.9.2 for detail). In particular, the *Data Governance Act*[2] (DGA) and the *Data Act*[3] are the two regulations that address horizontal (i.e., cross-sectoral) aspects of the strategy while a series of regulations on *data spaces* that address vertical (i.e., sector-specific) aspects. The European Commission has published a list of the envisioned *Common European Data Spaces*[4]. The first such data space is the *European Health Data Space* (EHDS) that, at the time of writing, exists as a Commission Proposal[5].

In data spaces, data that was collected for *primary use* is then shared at a large-scale for *secondary use*. In many data spaces, a significant portion of data in primary use is personal. This is evidently the case for health data in the *European Health Data Space* (EHDS). In this context, rendering the data usable means to find strategies of secondary use that comply with the requirements of data protection, in particular with the GDPR. This requirement is also clearly stated in the Data Governance Act that states: "*In the event of a conflict between this Regulation and Union law on the protection of personal data or national law adopted in accordance with such Union law, the relevant Union or national law on the protection of personal data shall prevail.*"[6].

Consequently, the work in Task 4.9.7 focuses on strategies for GDPR-compliant secondary use of personal data in the context defined by the *European Health Data Space*.

As stated in the *AnoMed* project plan, there are three main scenarios to consider:

(i) Secondary use of data that is clearly still personal,

(ii) secondary use of data that has been anonymized but where an undeniable residual risk of re-identification is present, and

(iii) secondary use of data that has been successfully anonymized such that the risk of re-identification is insignificant.

The present Task 4.9.7 focuses on scenarios (i) and (iii), while scenario (ii) is subject of Task 4.9.8.

In the context of the EHDS, the two scenarios that are in focus correspond to two kinds of secondary use (see Art. 44 EHDS and also Deliverable D4.9.2). In particular, they are:

- The use of **pseudonymous data** within a controlled ***secure processing environment*** (see Art. 2(20) DGA) by **vetted** *data users* authorized by a ***Data Permit*** (Art. 46 EHDS) that results from a ***Data Access Application*** (Art. 45 EHDS).
- The use of **successfully anonymized data** by arbitrary data users based on a **Data Request** (Art. 47 EHDS).

The present deliverable therefore explores strategies to implement the above two kinds of secondary use in a GDPR-compliant manner. In addition, the requirement of Art. 46(11) EHDS is taken into

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066
[2] https://eur-lex.europa.eu/eli/reg/2022/868/oj
[3] https://eur-lex.europa.eu/eli/reg/2023/2854/oj
[4] https://digital-strategy.ec.europa.eu/en/policies/data-spaces
[5] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0197
[6] See Art.

account, which states that the results of the former kind must be (successfully) anonymized in order to be published on the website of the involved data access body.

2 Data Protection Requirements for Data Spaces

Rendering personal data usable means to find strategies of secondary use that comply with the requirements of data protection, in particular with the GDPR. This section therefore focuses on data protection requirements imposed by the GDPR.

The most relevant requirements are expressed by the principles of Article 5 GDPR. Of these, *data minimization*, *purpose limitation*, and *storage limitation* are the most relevant for the discussed strategy options. These will therefore be discussed in more detail in the following.

The other principles, namely *lawfulness, fairness and transparency*, *accuracy*, *integrity and confidentiality*, and *accountability* are either already addressed by the legal act that lays the basis for secondary use (such as the EHDS) or only become a concern when implementing the discussed strategy options.

2.1 Data Minimization

The principle of *data minimization* is described in Art. 5(1)(c) GDPR. In particular, its wording is as follows: "Personal data shall be adequate, relevant and **limited to what is necessary in relation to the purposes** for which they are processed". (Emphasis added by author).

Data minimization not only concerns the information content but also has a temporal component[7]. In a setting where different processing phases are executed by different actors, the following requirements therefore apply:

> [1.1] Every actor shall have access to personal data only if that is necessary for the purposes.
>
> [1.2] If so, the information content and storage/access time shall be minimized to what is actually necessary.

These requirements have to be applied to the setting of secondary use in data spaces. This is illustrated in Figure 1. Here, personal data was originally collected for primary use by a *data holder*. This data is then made accessible to a *data user* for secondary use. The data user could have a direct relationship with the *data holder* or, more commonly, an indirect relationship through one or several intermediary *data access bodies*. The figure shows two intermediary *data access bodies*, which could for example be at national and European level.

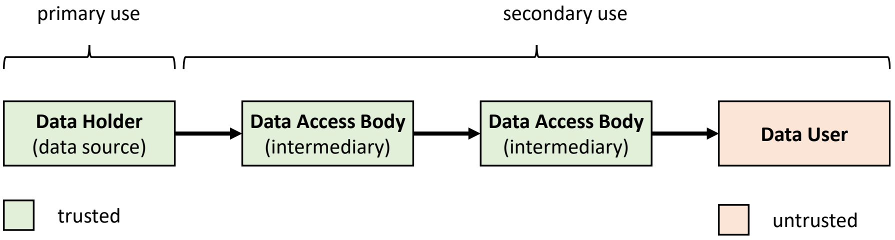

primary use
secondary use
Data Holder (data source)
Data Access Body (intermediary)
Data Access Body (intermediary)
Data User
trusted
untrusted

Figure 1: Secondary use of personal data in data spaces.

[7] See for example the EDPS Glossary entry for data minimization at https://www.edps.europa.eu/data-protection/data-protection/glossary/d_en#data_minimization (last visited 3/7/2024).

The figure uses color to indicate the level of trust assumed for the different actors in the setting. This is useful to illustrate that the *data minimization* requirement is independent of trust. In particular, any actor, whether trusted or not, must limit their access to personal data to what is actually required by the purposes[8].

The above discussion renders it evident that data access bodies simply accumulating copies of the primary use data would by no means be GDPR-compliant but much rather disregard the principle of *data minimization*. For compliance, **smarter technical solutions** that are discussed as strategy options later **are necessary**. This is further underlined by the fact that data spaces must also support European-level analyses and that a European central pool of data from all constituent data holders would be an extremely high data protection risk due to its unprecedented scale (as well as likely politically impossible).

2.2 Purpose Limitation

The principle of *purpose limitation* is described in Art. 5(1)(b) GDPR. In particular, its wording is as follows: "Personal data shall be collected for specified, explicit and legitimate purposes **and not further processed in a manner that is incompatible with those purposes**; […]". (Emphasis added by author).

There are two major measures to limit the risk of personal data being processed for other than the specified purposes:

- *confidentiality* and
- *compartmentalization*

Confidentiality restricts access to data. Without access, data can obviously not be processed for any purpose. To implement purpose limitation, access has to be restricted to only parties who require the data to fulfil the purposes (i.e., "need to know"). This prevents unauthorized parties to process the data for other purposes.

The set of data belonging to a single access decision is called *compartment*. Compartmentalization is the concept of separating data used for different purposes into distinct compartments. Obviously, if instead, the data were combined in a single compartment, parties would get unjustified access to data such that they could process it for other, illegitimate purposes. This is illustrated in Figure 2.

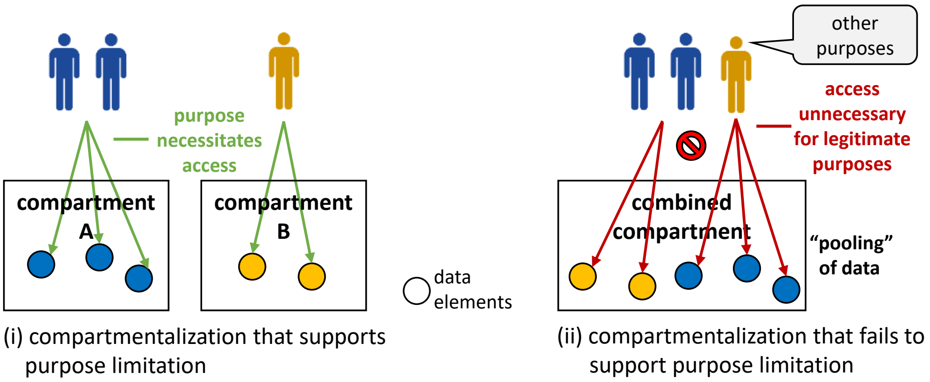


Figure 2: How compartmentalization can support purpose limitation.

[8] Note that Art. 34 EHDS lists the purposes that are legitimate for secondary use.

Note that parties with legitimate access to the data in a compartment could still process them for other purposes. Similarly, access control might fail (and cause a data breach). In these cases, like known from information security, compartmentalization is used to limit the possible risk.

Compartmentalization can be used both, horizontally (geographically) and vertically:

Horizontally, it means that it must be avoided to pool together data about different individuals originating from different sources. In other words, centralized national or even European "super pools" must be avoided.

Vertically, it means that it must be avoided to combine compartments from different sources that each contain data about different aspects of the same person. This vertical aspect is closely related to the concept of ***linkability*** (in the sense of database joins). If it is possible to combine (i.e., link) data of the same person across different compartments, obviously, the processing for a much wider range of purposes becomes possible.

Vertical compartmentalization is directly relevant for **pseudonymization** where a **pseudonym** is explicitly designed to limit (i.e., compartmentalize) to what extent data of the same person can be linked. In particular, unless data are part of a single pseudonymization, it is impossible to recognize and link data belonging to the same person. Vertical compartmentalization thus means that where possible, different pseudonyms (or pseudonym schemes) shall be used for different purposes.

In summary, the requirements of purpose limitation are the following:

[2.1] Where possible, data necessary for distinct purposes shall be organized into distinct compartments (with distinct access conditions).

[2.2] Unless necessary for the legitimate purposes, linkability of data pertaining to the same person across compartments shall be avoided. In the case of pseudonymization this means that the processing for distinct purposes shall use distinct, and thus unlinkable, pseudonyms.

Unsurprisingly, the combination of data from different compartments is an important indicator of elevated data protection risk. This is explicitly stated by the Article 29 Data Protection Working Party (the predecessor body of the European Data Protection Board) in its *Guidelines on Data Protection Impact Assessment (DPIA)*[9]. In particular, in the section[10] titled "*When is a DPIA mandatory? When processing is 'likely to result in a high risk'*", they provide nine **indicators for high data protection risk**. Number 6 reads: "**Matching or combining datasets**, for example originating from two or more data processing operations performed for different purposes and/or by different data controllers in a way that would exceed the reasonable expectations of the data subject".

2.3 Storage Limitation

The principle of *storage limitation* is described in Art. 5(1)(e) GDPR. In particular, its wording is as follows: "Personal data shall be **kept in a form which permits identification** of data subjects for no

[9] WP 248 rev.01, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, Adopted on 4 April 2017, As last Revised and Adopted on 4 October 2017, https://ec.europa.eu/newsroom/article29/items/611236 (last visited 3/7/2024).

[10] *Section III. B. a)*

longer than is necessary for the purposes for which the personal data are processed; personal data may be **stored for longer periods** insofar as the personal data will be processed solely for […], **research purposes** or statistical purposes **in accordance with Article 89(1)** subject to implementation of the **appropriate technical and organisational measures** […]". (Emphasis added by author). Article 89(1) GDPR further states that "[…] Those measures may include **pseudonymisation** […]". (Emphasis added by author).

While the principle's name "*storage limitation*" exclusively conveys the aspect of limiting the storage period, the wording of "**kept in a form which permits identification**" contains an important second aspect of the principle related to identification. The latter is even clearer considering the necessity of technical and organisatorial measures in accordance with Article 89(1). The only concrete measure that is actually named there is *pseudonymization*.

According to Art. 4(5) GDPR, pseudonymization is a manner of processing such "that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person". In other words, pseudonymization is a manner of processing in which technical and organisational measures protect against the actual identification of data subjects. These measures are cannot prevent every kind of identification[11], but at least drastically lower the risk of one occurring.

It is important to note that pseudonymization, while attempting to prevent identification, is substantially different from anonymization. In particular, identification is only prevented in the context of the technical and organizational measures that are implemented as part of the pseudonymization. Pseudonymous data are clearly still personal data[12].

In contrast, in the case of anonymous data, identification must not be possible independently of the presence of any technical organizational measures. Anonymous data therefore falls outside of the GDPR and can be freely published.

This leads to the following requirements for storage limitation:

[3.1] Personal data shall be **deleted or** (successfully) **anonymized** as soon as the purposes of processing allow it.

[3.2] Where the former is not possible, personal data shall be protected with technical and organizational measures to **minimize the likelihood of identification** of data subjects in accordance to what is necessary for the purposes.

The likelihood of identification is different for different types of pseudonymous data. This is illustrated in Table 1 that distinguishes three types of pseudonymous data. The requirement [3.2] thus mandates to choose the type of pseudonymity such as to limit the likelihood of identification of data subjects to the minimum that is required by the purposes.

[11] For example, pseudonymization cannot prevent "spontaneous recognition" where a legitimate actor possesses "additional information" in her head and simply recognizes a known data subject by its data.

[12] This is stated explicitly in the 2nd sentence of Recital 26 GDPR.

Table 1: Likelihood of identification for different types of data.

Relation to GDPR		Type of Data	Necessary Data Transformation	Effort/Difficulty	Identification
Personal data (inside GDPR)		**Directly identified**	none	none	100%
	Pseudonymous data	**Reversibly pseudonymous**	**Separate** & protect direct identifiers	Medium (TOMs to control identification)	Certain but controlled
		Irreversibly pseudonymous	**Remove** direct identifiers	Minimal (fewer TOMs)	Can happen unintentionally[13]
		Aggregated pseudonymous	**Aggregate values** over several persons	Minimal	Requires intentional re-identification[14]
Outside of GDPR	Anonymous data	**Truly anonymous**	Successfully **anonymize**	**Substantial**	Very unlikely
	--	**No data**	**Delete**	Minimal	Impossible

3 Data Spaces and Data Protection Risk

The previous section listed the major requirements imposed by the GDPR. This section discusses how stringent and with how much effort these requirements have to be fulfilled. It also may give insight which kind of processing is the most critical.

The GDPR is often seen as a risk-based approach. Here, the risks to the rights and freedoms of natural persons are mitigated by implementing measures in support of the data protection principles[15]. Such mitigation can vary in its degree. In particular, there is a continuum in the rigor, sophistication, and thus necessary effort and cost of choosing and implementing mitigation measures. High risk then requires more substantial mitigation than low risk. Where a high risk cannot be sufficiently mitigated, processing cannot even commence without the prior consultation of the competent data protection supervisory authority[16].

In this context, it is important to be able to recognize factors and areas of high risk. In particular, this can guide designers and implementers on where a more careful attention to GDPR-compliance is required.

There exists official guidance on how to recognize high risk. In particular, to determine whether a more detailed Data Protection Impact Assessment (DPIA) is required, in a first step, the overall risk

[13] Unintentional identification by actors with legitimate access can happen in the form of "spontaneous recognition". For a definition of the term, see for example page 30 of ESSNET SDC, Handbook on Statistical Disclosure Control, Version 1.2, https://cros.ec.europa.eu/system/files/2023-12/SDC_Handbook.pdf (last visited 4/7/24).

[14] A good example is provided by the *DataShield* project [https://www.datashield.org/] in their post [https://datashield.discourse.group/t/statement-datashield-disclosure-controls-and-mitigation/628] which in turn refers to the scientific paper https://doi.org/10.1101/2022.10.09.511497.

[15] See Article 25 GDPR.

[16] See Article 36 GDPR.

level has to be assessed[17]. For this purpose, the *Article 29 Data Protection Working Party* has issued official guidance[18] on how to determine this overall risk level. This guidance has later also been endorsed[19] by the *European Data Protection Board* (EDPB). It consists of a list of **nine criteria** that indicate particular risk. The guidance states that "[i]n most cases, a data controller can consider that **a processing meeting two criteria** would require a DPIA to be carried out". Then, the processing activity likely **possesses an inherent high risk**.

The following looks at the criteria that seem to apply to data spaces:

- *Criterion 5*: **Data processed on a large scale**:
  The guidance suggests that to determine scale, among others, "the number of data subjects concerned", "the volume of data […] being processed", and "the geographical extent" shall be considered. Data spaces implementing a European-wide large-scale data sharing, this criterion seems to apply to all data spaces in which personal data is shared.

- *Criterion 6:* **Matching or combining datasets**:
  The guidance suggests that to determine *matching or combining*, among others, "originating from two or more data processing operations performed […] by different data controllers. This is also clearly the case since the primary use processing operations by distinct primary use controllers are then made available in a combined manner for secondary use. This criterion thus also seems to apply to all data spaces in which personal data is shared.

- *Criterion 8*: **Innovative use or applying new technological or organizational solutions**: This criterion considers how much experience there is in handling and governing the solutions used for a processing activity. The following reasons that this criterion indeed applies to all **data spaces in which heavily rely on the efficacy of anonymization** as an enabling actor of data sharing. This is for example the case in the European Health Data Space where a large portion of secondary uses is enabled by anonymization[20]. There is currently insufficient experience concerning the efficacy of anonymization in a setting with a large number of related data disclosures that can erode the so-called privacy budget.

  In more detail, anonymization generally fails to reduce the risk of re-identification to zero[21]. Much rather, a residual risk of re-identification is always present. In the scientific literature, ample reporting has illustrated that all anonymization techniques that lack strong mathematical guarantees (such as *K-Anonymity*) are subject to re-identification attacks[22].

  Techniques that provide strong mathematical guarantees (such as *epsilon-Differential Privacy*)

[17] See Article 35 GDPR.

[18] WP 248 rev.01, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, Adopted on 4 April 2017, As last Revised and Adopted on 4 October 2017, https://ec.europa.eu/newsroom/article29/items/611236 (last visited 3/7/2024).

[19] See https://www.edpb.europa.eu/our-work-tools/general-guidance/endorsed-wp29-guidelines_en (last visited 5/7/24)

[20] See Article 44(2) EHDS Commission Proposal.

[21] See Recital 64 EHDS Commission Proposal.

[22] See for example Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. 2017. "Exposed! A Survey of Attacks on Private Data." Annual Review of Statistics and Its Application (2017). https://privacytools.seas.harvard.edu/sites/projects.iq.harvard.edu/files/privacytools/files/pdf_02.pdf and Henriksen-Bulmer, Jane & Jeary, Sheridan. (2016). Re-identification attacks—A systematic literature review. International Journal of Information Management. 36. 1184-1192. 10.1016/j.ijinfomgt.2016.08.002. http://dx.doi.org/10.1016/j.ijinfomgt.2016.08.002 (both last visited 9/7/24).

are resistant to such attacks, but only under the condition that the so called *privacy budget* is not eroded. Here, every (statistical, anonymized) data that is released about the same characteristics "leaks" information about its data subjects. The privacy budget can then be eroded by combining the leaked information across a multitude of such data releases. In other words, the leaked information can be accumulated to re-identify data subjects.

The practical risk of re-identification and how to manage the privacy budget across large numbers of releases is still little understood. In this setting, anonymization must therefore be considered a "new technology". This is for example evident when considering that the U.S. Census Bureau conducted the likely first large-scale practical re-identification attack and was surprised by its unexpected effectiveness[23]. Reliance on a little understood safeguard obviously raises the risk.

In addition, data spaces even have some properties that facilitate re-identification attacks based on a multitude of releases:

  - Data releases are cataloged and thus can be easily discovered by attackers with automated procedures.
  - The data in a data space are harmonized and releases can thus easily combined in a reconstruction attack.
  - The possibility of unlimited data requests[24] and mandatory centralized publication of certain releases[25] renders the availability large numbers of releases highly likely.

- *Criteria 4 & 7:* **Sensitive Data** & **Data concerning vulnerable data subjects:**
  For the first criterion, the guidance suggests to take *special categories of personal data* which are the subject of Article 9 GDPR into account. These criteria likely apply in certain data spaces. For example, health data shared for secondary use in the European Health Data Space clearly fall into special categories of personal data and patients are typically considered to be vulnerable.

Considering that not only two, but at least three and, depending on the data space, up to 5 criteria apply, the overall data protection risk must definitely be considered to be high. A very careful implementation of mitigating measures is therefore required, even in the case that this results in a substantial cost.

While most criteria apply due to the base characteristic of a data space, *matching and combining* can be used as an indicator of what functionality or architectural area poses particular risks and thus require particularly careful attention.

[23] Garfinkel, Simson & Abowd, John & Martindale, Christian. (2019). Understanding database reconstruction attacks on public data--These attacks on statistical databases are no longer a theoretical danger. Communications of the ACM. 62. http://dx.doi.org/10.1145/3287287 or https://queue.acm.org/detail.cfm?id=3295691 (last visited 9/7/24).

[24] See Article 47 EHDS Commission Proposal.

[25] See Article 46(11) EHDS Commission Proposal.

4 Main Characteristics of Secondary Use

Data spaced can comprise a diversity of use cases that require different implementations and measures to comply with data protection requirements. Covering all possible cases is beyond the scope of the present deliverable. To capture the diversity partly, two characteristics of use cases are discussed here, namely:

- The level of identification that is required by a use case, and
- whether and how the necessary data has to be integrated from multiple sources.

A detailed description is given in the following subsections.

4.1 Required level of identification

The present subsection focuses on the level of identification that is necessary to fulfil the stated purposes of a use case. In particular, it proposes a taxonomy of different levels of identification.

Regulations such as the European Health Data Space[26] differentiate between three levels of identification:

- **Directly identified** data (confined to primary use)
- **Pseudonymous** data[27] (where necessary for the purposes of secondary use)
- **Anonymous** data[28] (where compatible with the purposes of secondary use)

In line with sub-section 2.3 on the principle of *storage lim*itation, the following proposes that **a more diversified taxonomy** is useful for the understanding of use cases.

The basic concept of levels of identification states that the less identification is possible, the more the possible uses are restricted. This is well-known from the discussion of the **trade-off between level of anonymity and utility**[29].

[26] In its Commission Proposal.

[27] See Article 44(3) EHDS Commission Proposal.

[28] See Article 44(2) EHDS Commission Proposal.

[29] See for example: Domingo-Ferrer, J., Ricci, S., Soria-Comas, J. (2017). A Methodology to Compare Anonymization Methods Regarding Their Risk-Utility Trade-off. In: Torra, V., Narukawa, Y., Honda, A., Inoue, S. (eds) Modeling Decisions for Artificial Intelligence. MDAI 2017. Lecture Notes in Computer Science, vol 10571. Springer, Cham. https://doi.org/10.1007/978-3-319-67422-3_12,
Grigorios Loukides and Jianhua Shao. 2008. Data utility and privacy protection trade-off in k-anonymisation. In Proceedings of the 2008 international workshop on Privacy and anonymity in information society (PAIS '08). Association for Computing Machinery, New York, NY, USA, 36–45. https://doi.org/10.1145/1379287.1379296,
Gu Yonghao and Wu Weiming, "A quantifying method for trade-off between privacy and utility," IET International Conference on Information and Communications Technologies (IETICT 2013), Beijing, China, 2013, pp. 270-273, doi: 10.1049/cp.2013.0062.

The proposed taxonomy of *levels of identification* is represented in the following table:

Table 2: Taxonomy of levels of identification.

Level of identification	**Identification Risk**	**Utility** (examples)
directly identified	All data subject are identified.	Baseline without restrictions (necessary for primary use).
reversibly pseudonymous	Only few selected data subjects are re-identified in special cases.	• Contacting data subjects possible under specified conditions and with approval. • Additional data about selected data subjects can be requested.
irreversibly pseudonymous	Identification is unlikely but can happen unintentionally.	• Data subject can no longer be contacted. • No additional information for a given data subject can be requested.
aggregated pseudonymous	Identification is unlikely and possible only with an explicit effort given that appropriate measures are implemented.	• Data of individuals not accessible. This can for example hinder the assessment of data quality. The input data is complete (includes outliers) and truthful, however.
(successfully) anonymized	Highly unlikely even in absence of measures.	Uncertainty of analysis result since data is: • less detailed (generalization), • a subset (eliminate outliers), or • not truthful (protected by noise).

The taxonomy is described in more detail in the following:

Directly identified:
Direct identification is typically **necessary for the purposes of primary use** only and not permissible in secondary use. At this level, identification of data subjects is unrestricted. All data subjects are identified. This level represents the base line without any restrictions of utility.

Reversibly pseudonymous:
Here, pseudonymization is designed to prevent identification of data subjects during ordinary processing. In special cases, identification (i.e., reversal of the data pseudonymization) is still possible. It is typically only possible under well-defined circumstances and requires approval. Consequently, most data subjects remain unidentified; only a selected few are (re-)identified when special conditions justify it. Since the data is still pertains to individual, it is possible that (even without direct identifiers), data users can recognize data subjects they know based solely on their data. This is called “spontaneous recognition”[30].

[30] For a definition of *spontaneous recognition*, see for example page 30 of ESSNET SDC, Handbook on Statistical Disclosure Control, Version 1.2, https://cros.ec.europa.eu/system/files/2023-12/SDC_Handbook.pdf (last visited 4/7/24).

The utility at this level is basically unrestricted since critical uses (most prominently identification) are not rendered impossible but solely controlled. This permits in particular to:

- **contact data subjects** when the secondary use produces results that are in their interest to know, and
- break out of the compartment of the pseudonym domain by **requesting**[31] **additional data** about a selected data subject. Such a "zooming in" on a person can be motivated by a follow-up to **verify a hypothesis** or as part of **exploratory analysis**.

Irreversibly pseudonymous:
Irreversible pseudonymization renders it impossible to break out of the compartment of the pseudonym domain even in special cases. Since the data still pertains to individuals, unintentional *spontaneous recognition* still remains a risk. Consequently, fewer data subjects are identified compared to the preceding level of identification.

The utility that comes with a planned, but controlled, reversal of the data pseudonymization falls away at this level. But in consequence, it can significantly reduce the necessary effort for complying with the GDPR. In particular, it eliminates the need for the typically technical effort to protect access to the re-identifying information, and the typically organizational effort to process reversal requests. In addition, the data protection risk of unauthorized re-identification[32] falls away.

Aggregated pseudonymous:
At this level of identification, pseudonymous data are only accessible in aggregated form. This could for example take the form of statistics where each data value is calculated from the individual values of a minimal number of data subjects. Another example for such aggregation is a model resulting from machine learning trained by the individual data of many data subjects.

In these cases, identification is still possible (and the data is therefore considered to be pseudonymous) but only if an explicit effort to re-identify data is made. An example of such an intentional re-identification for statistical data has for example been documented by the DataShield project[33].

Since re-identification is now only possible through explicit (and evidently malicious) action, the likelihood of data subjects being identified can be highly reduced. This relies on the implementation of **adequate technical and organizational measures**, however. An example of an organizational measure is the prohibition of re-identification efforts[34]; and example of a technical measure is the supervision of a secure processing environment[35] through logging and anomaly detection of processing operations. Together, such measures can successfully prevent identification.

Without access to individual-level pseudonymous data, the utility is reduced compared to the previous level of identification. This includes the difficulty of assessing data quality based only on aggregate data (e.g., to recognize errors in the data). But for a multitude of purposes, the utility is by

[31] Such a request is obviously subject to approval.

[32] Such unauthorized re-identification could be attempted either by insiders with legitimate access or by external attackers.

[33] *DataShield* project [https://www.datashield.org/], post [https://datashield.discourse.group/t/statement-datashield-disclosure-controls-and-mitigation/628] which in turn refers to the scientific paper https://doi.org/10.1101/2022.10.09.511497.

[34] See Article 44(3) EHDS Commission Proposal.

[35] See Article 2(20) DGA.

all means still sufficient. In particular, the aggregate data values (such as statistics) are based on truthful (i.e., not altered by noise) and complete (i.e., including outliers and data of "outstanding" subjects) data sets. This provides certainty about the correctness of results (such as the significance level of statistical tests). In other words, results remain unaffected of possible artefacts and biases caused by anonymization.

Successfully anonymized:
At this level of identification, the identification of data subjects is highly unlikely even in absence of technical or organizational measures.

The utility of the data is again reduced in comparison of the previous level of identification. In particular, effects of anonymization such as noise addition, outlier elimination, or generalization could render it impossible to detect more subtle patterns in the data. Anonymized data may leave a doubt whether less clear-cut results are indeed significant or instead an artefact of the anonymization.

Data at this level are ideally suited to provide data to a large number of data users without the need of formal vetting and supervision of applicants or evaluation and approval of applications. In this setting, in cases where results leave doubts, a follow-up request for data at a higher level of identification always remains possible[36]. An initial study that is based on anonymized data may also be a highly convincing justification in an application for more detailed access to data at a higher level of identification. Note that the EHDS concepts of un-bureaucratic "data requests[37]" of anonymous data and more onerous "data access applications[38]" for pseudonymous data fit very well with this strategy.

The level of identification that is necessary for the purposes pursued by a given data user seems to be a very important characteristic for distinguishing use cases. The provided taxonomy that defines the possible levels of identification will be use later to generate use cases.

The key distinction is certainly between pseudonymous and anonymous data. Anonymization typically has the following characteristics:

- It reduced detail (e.g., through generalization);
- it removes outliers and data of exceptional data subjects that do not blend into a group of similar data;
- it adds deviations from "true" values (e.g., the noise used in epsilon-Differential Privacy); and
- anonymized data better support certain types of analysis than others. [XX verbally: UHL, literature here].

In contrast, pseudonymous data are truthful, contain full detail, include data of exceptional data subjects, and are suited for a wider range of analysis types. These properties may justify the use of pseudonymous data since the purposes of a given study may be unreachable when using anonymous data.

[36] Note that also so-called "verification servers" that are discussed later in this document could be used to correctly assess results that are based on anonymized data.
[37] See Article 47 EHDS Commission Proposal.
[38] See Article 45 EHDS Commission Proposal.

4.2 Integration of partitioned data

The second characteristic of use cases is how they assemble their data from different sources. If the required data come from different sources, the term *partitioned*[39] *data space* is used. Partitioned data spaces then require the *integration* of the physically distributed data set into a single logical one. This subsection describes the possible cases of such integration.

Data spaces typically comprise a multitude of data sources (represented by *data holders*). Data spaces typically aim to ease the administrative and technical burden of data users to piece together a puzzle from multiple data sources. The joint (integrated) analysis of data from multiple sources bears the potential of conducting analyses and gaining insights that would be impossible with just a single source or prohibitively onerous without the support offered by the data space.

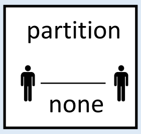
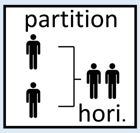
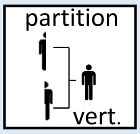
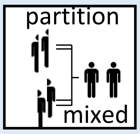
Data spaces partly ease the burden of data users administratively though harmonization of rules, metadata, and the possibility to request (apply for) data of multiple sources in a single step. In addition, it employs intermediaries (such as *data access bodies*) to ease the burden through technical means.

The main distinction of use cases lies in how data need to be logically[40] combined to support a particular analysis by a data user In the context of personal data, there four types of partition that represent possible logical combinations. Table 3 Table 1illustrates them.

[39] "In mathematics, a **partition of a set** is a grouping of its elements into non-empty subsets, in such a way that every element is included in exactly one subset. https://en.wikipedia.org/wiki/Partition_of_a_set (last visited 11/7/24).

[40] Note that the term "logical" was expressly used to indicate that the combination may not necessarily be physical. Generally, there are multiple physical options of how to implement a logical combination.

Table 3: Four types of partition possible in data spaces.

type of partition	**definition**	**example**	**visualization**
none	Data comes from a single source.	A single source provides a data set of person's height. The analysis computes and average height.	
horizontal	Multiple data sources provide the same attributes about different persons.	A multitude of sources provide data sets of height of persons living in their geographic area of operation. The analysis computes the overall average height over the combined geographic area.	
vertical	Multiple data sources provide different attributes about the same persons.	To compute the average Body Mass Index (BMI)[41], two data sources must be combined: one providing height data of a given population; the other providing weight data of the same population.	
mixed	Multiple data sources provide different attributes about different persons.	The prevalence of a certain disease in Europe is estimated by geographic, horizontal integration across healthcare providers. A vertical integration is necessary at least at a regional level to avoid the possibility of duplicates, where patients visited multiple providers.	

Due to the scale and conception of data spaces, use cases that requite only a **single data source** (and thus avoid the need of integrating partitioned data) are likely rare. Note that an exception may be the access to confidential statistical microdata which is regulated by Commission Regulation (EU) No 557/2013 and is described in recital 7 DGA as an important experience with secure processing environments at European level. Here, the data sources seems to be exclusively national statistics authorities[42] who already are in possession of nationally integrated micro data. An integration of microdata at European level does not seem to be foreseen.

Horizontal integration is likely required by a majority of use cases. This is evident in the fact that in many data spaces, data sources can be expected to be regional or even local. Since many use cases

[41] The BMI is defined as the ratio of the weight and the square of the height.

[42] For example, the „*access facility*", i.e., the *secure processing environment* (in DGA terminology), is located within national statistics authorities (see Art. 8(2) Commission Regulation (EU) No 557/2013).

can be expected to have a wider geographic scope, horizontal integration across composite geographic units is required. Figure 1Figure 3 illustrates this.

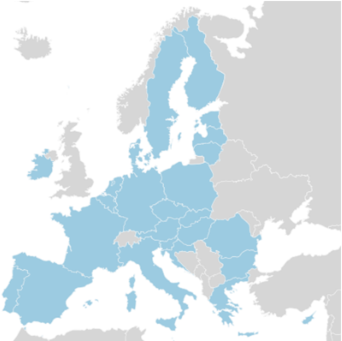


Figure 3: Data in Europe are typically partitioned horizontally (geographically) by political/administrative boundaries.

The ability of **vertical integration**, i.e., to combine data about different aspects of life, is often seen as an important factor for exploiting the full potential of a data space. In the EHDS proposal, it is unlikely that regarding a given person, the foreseen data categories (see Art. 33) can be collected by a single data source, such that vertical integration would be unnecessary. For example, actual health data (i.e., EHR in Art. 1(a) EHDS) and "data impacting on health" (Art. 1(b) EHDS) for a given person are unlikely to be collected by a single actor. While the EHDS proposal does not seem to address vertical integration explicitly, the need to vertically integrate health data with factors that impact on health seems to be a likely implication.

The German law on rendering health data usable for research purposes[43] explicitly addresses a case of vertical integration in its paragraph 4. The objectives of a potential German law on research data, as described by the German Federal Ministry of Education and Research[44], explicitly include the need for vertical integration in its Section 4, second paragraph. One aim of the potential legislation is to provide legal certainty and provide the necessary legal basis in this area. Finally, vertical integration also already seems to be used in research practice[45].

Mixed integration which combines horizontal and vertical integration is the most general case. It can be expected that use case require a mixed integration across data sources.

[43] Gesundheitsdatennutzungsgesetz – GDNG, § 4 Verknüpfung von Daten des Forschungsdatenzentrums Gesundheit mit Daten der klinischen Krebsregister der Länder, https://www.gesetze-im-internet.de/gdng/__4.html (last visited 24/7/24).

[44] Planned Forschungsdatengesetzes (FDG), as described in *Eckpunkte BMBF Forschungsdatengesetz*, Stand: 28.02.2024, https://www.bmbf.de/SharedDocs/Downloads/de/2024/240306_eckpunktepapier-forschungsdaten.pdf (last visited 24/7/24), section 4, 2nd paragraph.

[45] See for example https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set (last visited 24/7/24).

5 Selected Measures for Data Spaces

This section discusses selected technical and organizational measures that can be implemented in data spaces in support of data protection requirements. No attempt is made to present a comprehensive list of all necessary measures nor is it implied that a described measure is sufficiently effective to satisfy a certain data protection principle.

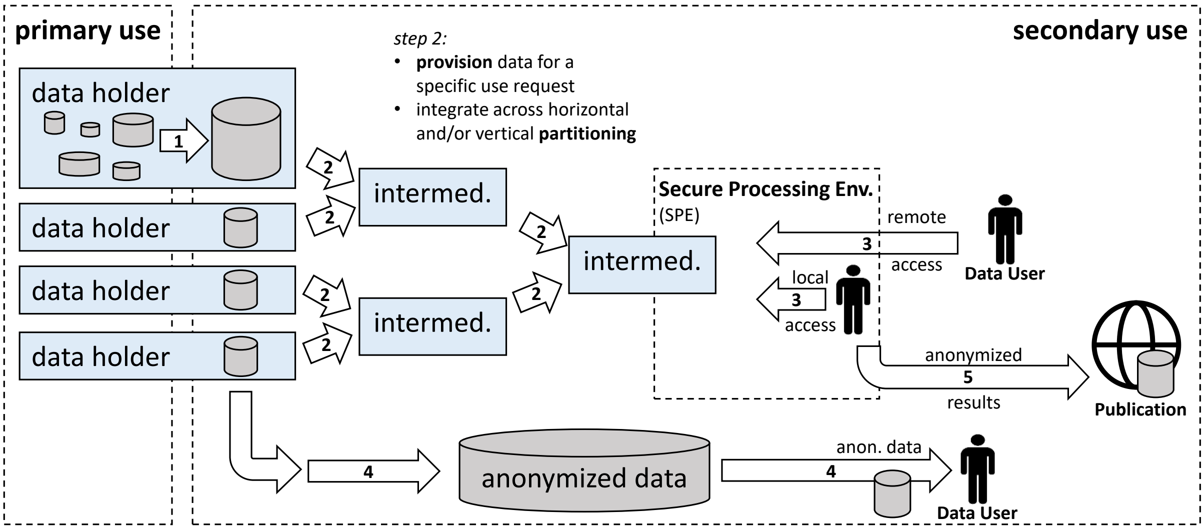


Figure 4: Organization of discussed measures.

The discussion of selected measures is organized as indicated in Figure 4. In particular, the figure distinguishes different processing steps that are indicated by numbered arrows. The description of measures is then organized according to the step in which a measure is applicable.

In more detail, the steps are the following:

1. Provisioning primary use data for secondary use;
2. provisioning the required data for a specific data use across multiple data sources and intermediaries[46] (e.g., in response to a *data access application* in the EHDS[47]);
3. providing access to pseudonymous data within a secure processing environment;
4. providing anonymized data to data users (i.e., a *data request* in the EHDS[48]); and
5. publishing anonymized (or anonymous) results of analysis of pseudonymous data[49].

Steps 4 and 5 are assigned in this order since the discussion of step 5 refers to arguments used in step 4.

[46] In the EHDS proposal, such intermediaries correspond to health data access bodies (see Art. 36 EHDS proposal).

[47] See Art. 45 EHDS proposal for data access applications, see Art. 46(3) EHDS on provisioning.

[48] See Art. 47 EHDS proposal.

[49] See Art. 46(11) EHDS proposal.

5.1 Step 1: Provisioning primary use data for secondary use

This section discusses data protection of the processing aspect of provisioning data for secondary use.

5.1.1 Setting

The following describes the setting and assumptions of the discussion.

In data spaces, (personal) data that is collected and processed for the purposes of primary use shall be made available also for secondary use. The following discussion assumes that primary use data are managed in a heterogeneous collection of databases and that their use is often mission critical.

The primary use (personal) data continually evolve over time. In particular, data of new data subjects are typically added to databases and additional data about already existing data subjects are collected.

Further, data spaces encompass a multitude of holders[50] of primary data (*data sources*) who also may collect data about the same persons (i.e. their sets of data subjects may overlap). This is for example the case in healthcare where people turn to different healthcare providers for different kinds of health issues or for a second opinion concerning the same issue.

The data provided for secondary use is a subset (and possibly a derivate) of the primary use data.

It is also assumed that in primary use, a long-term[51] person identifier is available that is consistently used to identify a person.

Possible optional requirements for secondary use may contribute to the setting. Some of the relevant optional requirements are listed in the following:

- ***Pseudonymization reversal***: Here, it must be possible to obtain the primary use identity of a person that was singled out during secondary use. Assuming that pseudonyms are used to refer to persons in secondary use, it must thus be possible to obtain the full identity of the primary use (i.e., the long-term identifier) from the pseudonym. This will be referred to as ***pseudonymization reversal***. An example of a data space where this is a requirement is the EHDS[52]. This requirement implies that a long-term **primary use person identifier** is available and it is related to the pseudonym(s) employed in secondary use.

- ***Vertical integration***: Where data about a given person are spread across multiple data sources, their logical integration must be possible. This requirement will be referred to as ***vertical integration***. (See also Section 4.2). During primary use, vertical integration is enabled by the long-term person identifier; in secondary use, this requirement determines the compartmentalization that is possible through the creation of different pseudonyms for different uses.

- ***Reproducibility***: Here, an analysis conducted in the data space must be **reproducible**. Reproducibility is an important concept of the EU's research and innovation strategy. In

[50] The term "data holder" is defined in Article 2(8) DGA.

[51] Note that "long-term" may be shorter than "life time". For many persons, it may indeed mean "life-time" but it is not excluded that in certain cases, it changes. For example, an Italian Codice Fiscale may change when the name is changed (note: marriage does not require a name change of Italian women but can for foreign residents); a German insurance number can change when a person changes to another insurer. The difference between long-term and life-time can likely be practically neglected for secondary use.

[52] See commission proposal, Articles 38(3) and 44(3).

particular, one of the five ***open science** practices*[53] is "ensuring verifiability and reproducibility of research outputs". These practices also include the ***FAIR*** principles (Findable, Accessible, Interoperable, and Reusable). FAIR's principle F1 reads "(Meta) **data** are assigned **globally unique and persistent identifiers**"[54]. Technically, such identifiers typically take the form of Universal Resource Identifiers (**URI**s)[55]. Obviously, for reproducibility, a given URI must consistently identify the very same data. This is particularly relevant for data that continuously evolves over time.

- ***Long-Term Linkability*:** Data of a given person that are collected at different points in time must be linkable. This optional requirement is for example important in medical long-term studies where the outcome of treatment is evaluated over several years.

To provision data for secondary use in this setting, three basis architectural options are available that are discussed in the following subsection.

5.1.2 Architectural options

The three available options are:

(i) Grant access to primary use databases,

(ii) Create distinct secondary use databases that constitute "filtered" mirrors of the primary use ones that track the evolution (in time) of data, and

(iii) Create distinct secondary use data bases that refrain from tracking the evolution of data but represent only selected snapshots of this evolution.

Option (i) is problematic for a multitude of reasons including the following:

- Often mission critical database operations are exposed to an additional, possibly unpredictable volume of queries.
- The personnel managing the primary use data bases is charged with additional tasks and responsibilities.
- Even a small glitch in the configuration of access conditions could result in unauthorized access (including modification or deletion) of primary use data.

Option (ii) is also problematic for several reasons:

- The topology of databases used in the primary use domain is likely different from that in the domain of secondary use. Also, the database technologies may be diverse. This renders any mirroring difficult.
- While mirroring is well-suited for constructing subsets of data, more computation-intensive and time consuming derivations such as anonymization are likely problematic.

[53] https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en (last visited 23/7/24).

[54] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18, Box 2: The FAIR Guiding Principles, page 4.

[55] See IETF RFC 3986.

- In case that reproducibility for the analysis of secondary data is required, mirroring results in the "most current state" of data while reproducibility requires the state of data at a given point in time. Consequently, a data base of a single state would have to be mapped to a "historic database" that incorporates version control and tracks changes. This may result to be highly complex.

Due to these difficulties with the above two options, only the **architecture of option (iii)** is discussed in further detail. The architectural option is illustrated in Figure 5.

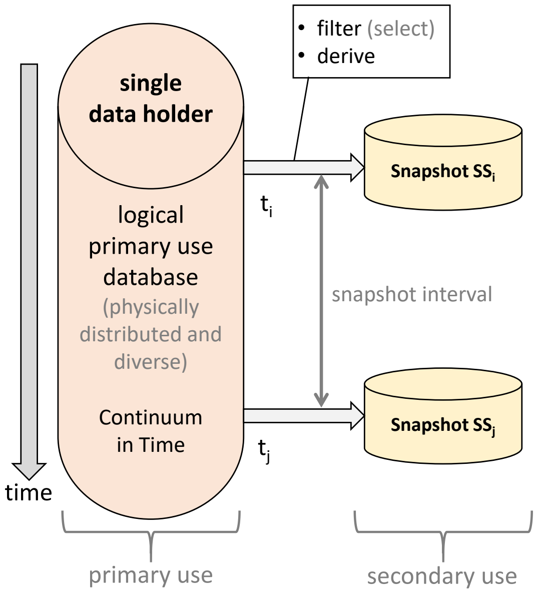


Figure 5: Architectural option based on snapshots.

The figure shows how the secondary use data are represented by snapshots of the primary use data. The architecture's scope is limited to a single data holders (possibly comprising multiple internal sources) but otherwise can encompass a **distributed topology of databases** and a **diversity of database technologies** on both sides (i.e., the primary as well as the secondary use side).

Snapshots are typically taken in regular **intervals**.

The snapshots typically only contain a **subset** of the primary use data. Subsets can be created for example by filtering the types of data (i.e., columns in a table) or exclude specifically sensitive "cases" (i.e., rows in a table).

Further, the snapshotting mechanism typically involves some kind of **derivation** such as pseudonymization, generalization (e.g., intervals instead of precise values). Snapshotting (in contrast to mirroring) permits to employ time-consuming derivations or even manual intervention. It further permits to employ derivations that cannot be used in an incremental or continuous manner (such as certain kinds of anonymization).

In this setting, a **snapshot** can then **identified** by the tuple formed by the identifier of the **data holder** (i.e., the DataHolderID) and the **time specification** (such as a date). A (logical) snapshot can contain a wealth of different data. It is assumed that for every data source, a **catalog** exists that lists all available information types using a standardized **information type identifier** (hence fore referred to as **catalogID**). So in summary:

DataSetID = DataHolderID + TimeSpec + ListOf(CatalogIDs)

Such a DataSetID is crucial for being able to specify data in a request. Since the ID also identifies versions in time, this snapshot-based ID is also instrumental for data spaces where **reproducibility** is required. In this case, it is important that snapshot are “frozen” and not allowed to change. The DataSetID is then a **Universal Resource Locator** (URI) as is required by the FAIR principles.

5.1.3 Data protection assessment and suitable measures

The following assesses architectural option (iii) from a point of view of data protection. For this purpose, it focuses on the data protection principles that were introduce in Section 2 on requirements plus selected additional considerations. It focuses on discussing relevant technical measures.

5.1.3.1 Data Minimization

According to the principle of *data minimization*, only the data actually necessary for the purposes must be provisioned for secondary use. Data spaces typically provide a list of admissible purposes. For example, this is the case in Article 34 EHDS (Commission Proposal). These purposes are stated very widely, for example “scientific research” (see Article 34(1)(e) EHDS Commission Proposal). To determine what data is necessary, more detailed assumptions about the purposes pursued by acceptable[56] data uses (by “data users”) may have to be made.

Data minimization thus requires explicit and reasoned decisions about the following questions:

- Which subset of data shall be provisioned for secondary use?
- Which level of detail is adequate for that data?
- Are there particularly sensitive cases[57] (or sub-categories of data) that need to be excluded or otherwise treated differently?

5.1.3.2 Storage Limitation

Section 2.3 has shown that the principle of *storage limitatio*n requires to minimize the likelihood of identification of data subjects. Since the purposes of secondary use do not require full identification of data subjects[58], fully identified data is inadmissible for secondary use which leaves only the possibility of **pseudonymous** or **anonymous data**. This is for example clearly stated in Article 44 EHDS Commission Proposal.

As was shown in Section 2.3, pseudonymous can be further separated in three different kinds of pseudonymous. The provisioning has to consider the maximally identified type of pseudonymous that is required in the data space; less identified forms can then easily be derived on the fly since the required derivation can be computed quickly and inexpensively.

[56] Acceptable in the sense that typically, a given concrete use of data is subject to approval, as for example through *data access applications* of Article 45 EHDS Commission Proposal.

[57] Note that Recital 64 EHDS Commission Proposal states such a special case. Further, Article 5(13) DGA addresses the special treatment of “highly sensitive” (but non-personal) “data categories”.

[58] Note that there may still be a possibility to re-identify selected data subjects under specific conditions. See for example Articles 38(3) and 44(3) EHDS Commission Proposal.

In case where data holders also have to provide anonymous data[59] to the data space, it must be considered that anonymization cannot be fast and inexpensive and is therefore ill-suited to be computed on the fly. It is therefore necessary to create anonymous data as part of the provisioning[60]. If anonymous data was not already available and cannot be created with appropriate response times, requests for anonymous secondary use data would either have to fail or provide pseudonymous instead of anonymous data. The former would put the functioning of the data space in question; the latter would violate the principle of storage limitation.

To satisfy the requirements of storage limitation, the **provisioning** of personal data for secondary use thus needs to **encompass both**, **pseudonymization and anonymization**.

Here, creating a pseudonymous snapshot is typically used as input for anonymization. The *DataSetId* (URI) for pseudonymous and anonymous data can thus share their main portion. In detail, the *DataSetID* then becomes the following:

DataSetID = IdentificationLevel + DataHolderID + TimeSpec + ListOf(CatalogIDs)

Where *IdentificationLevel* can be either *pseudonymous* or *anonymous*.

In this approach, the pseudonymous data set which corresponds to a given anonymous data set or vice versa can easily be identified.

5.1.3.3 Purpose Limitation

As illustrated in section 2.2, purpose limitation is closely related to compartmentalization. Here, the family of primary and that of secondary purposes must be separated into distinct compartments. The overall compartment of secondary use then supports a multitude of concrete secondary purposes that are pursued by actual data users[61] and can be separated in different compartments in later steps.

Compartmentalization can be achieved by preventing the linkability of data pertaining to the same person based on their identifier across distinct compartments. This requires an adequate design of pseudonymization of the snapshots. In particular, the design must compartmentalize such that

[59] Anonymous data is for example required for "data requests" of Article 47 EHDS Commission Proposal and can be requested directly from the data source according to Article 49 EHDS Commission Proposal.

[60] Note that logically, anonymization needs to be done as early as possible and thus by the data source. Organizationally, it may well be the case that some data sources lack the capacity to successfully anonymize and must therefore rely on another trusted actors (such as a *health data access body* in the EHDS) in the data space to provide the necessary support. Similarly, the data volume of a single data source may be insufficient for effective anonymization. In that case, it may be necessary that a trusted player, employing adequate protective measures, has to logically accumulate data from multiple sources to yield the necessary input volume for effective anonymization. An example of an adequate measure is distributed learning where data from the sources logically contributes to an anonymized AI model, but where no source data is physically transferred.

[61] The term "data user" is defined in Article 8(9) DGA.

secondary use data has the prime objective to prevent uncontrolled linkage[62] of primary and secondary use data is prevented.

Concretely, establishing a relation[63] between person identifiers in primary use data and pseudonymous identifiers in secondary use data should be possible only in the context of controlled pseudonymization reversal. It should only happen under well-specified conditions and be accessible only to specifically authorized personnel.

A secondary objective of the pseudonymization of secondary use data is to impose an upper bound on the linkability across secondary uses. This can be achieved by already using multiple compartments in the provisioned secondary use data. This is for example the case where compartments already exist in primary use and where it is decided that to avoid unnecessary risk, also in secondary use, the corresponding data shall not be linkable. Such compartmentalization could be motivated by risk reduction. Note that according to the Article 29 Data Protection Working Party, linking (i.e., matching) represents an indicator of high risk (see Section 3).

In this step, the principle of purpose limitation thus leads to the provisioning of data in one or several overall *secondary use compartments.* From here, data can be provided for concrete purposes pursued by data users. For concrete purposes, the compartmentalization can then be further refined (see later). This prevents for example that distinct data users can link "their" data across purposes and thus learn more about data subjects than necessary. In general, purpose limitation requires that as much compartmentalization is implemented as early as possible[64] for the purposes. Therefore, the first compartmentalization is already required during the provisioning of secondary use data in step 1.

Table 4 shows different scenarios of compartmentalization during this provisioning step 1. It illustrates different options for choosing a *pseudonym domain*, i.e., the extent where the same pseudonym is used for a given person. Within such a domain, data belonging to a given person can be linked.

[62] Note that the linkage based on data values can usually not be avoided.

[63] Note that this relation could be traversed in either direction.

[64] Note that Recital 49 of the compromise text of the EHDS states "Taking into account the specific purposes of the processing, data should be anonymised or pseudonymised as early as possible in the chain of making data available for secondary use." (see https://www.consilium.europa.eu/media/70909/st07553-en24.pdf, last visited 12/12/24).

Table 4: Different linking options (i.e. compartmentalization) for secondary use data. Here, ✓ *indicates the possibility to link,* ⊗ *the use of distinct pseudonym domains that prevent linking.*

	different information categories inside single data holder	**All data holders inside a certain group**	**all data holders of data space**
Pseudonym domain is subset of single data holder	⊗	⊗	⊗
Pseudonym domain is data holder	✓	⊗	⊗
Pseudonym domain are groups of data holders	✓	✓	⊗
Pseudonym domain is whole data space	✓	✓	✓

The top row shows the case where a data holder encompasses several independent internal sources and where it is unnecessary to link between them. Here, different information categories use different pseudonym schemes and thus represent distinct, unlinkable compartments. Considering that linking information across categories is a considered an indicator of high data protection risk (see Section 3), this is the most data protection friendly scenario. If linking goes beyond this scenario, it should be demonstrable that this is indeed necessary for the purposes.

The second row shows the case where all data of a given holder falls in the same pseudonym domain and is thus linkable. Data originating from different data holders cannot be linked, however. Note that a single data holder is responsible for the pseudonymization of data which renders the management of the necessary secret reversal information easier.

The third row shows how groups of data holders in the data space jointly use the same pseudonym domains. Thus, data from different holders in the same group can be linked. Linking across data holders can also be necessary for *vertical integration* (see Section 4.2). Since here, the pseudonymization domain encompasses multiple data holders, the secret reversal information needs to be shared.

From a data protection point of view, the smaller the groups are (i.e., the smaller the compartments), the better. Such groups could be defined thematically based on categories of information, or geographically, grouping data holders from the same administrative unit (such as region, province, or state). A geographic grouping is often reasonable when it is unlikely that a person's data cross certain kinds of boundaries. This is mostly the case with national boundaries between Member States (while data spaces are typically European). Therefore, pseudonym domains for groups are very common in data spaces. This also renders it unnecessary to coordinate and share pseudonym reversal secrets across nation boundaries (and thus jurisdictions).

The fourth table row illustrates the case where only a single pseudonym domain is used across all data holders of the data space. This is likely uncommon and it may prove difficult to justify the necessity of Europe-wide data linking, sharing of a secret, and possibility of pseudonym reversal.

While the kinds of compartmentalization discussed so far lie in the thematic and/or geographical dimension, compartmentalization is also possible in the temporal dimension. This is illustrated in Table 5.

Table 5: Compartmentalization in the temporal domain.

	Disjoint data collection periods
Long-term linkability necessary for purposes	Same person is assigned **same pseudonym**
Long-term linkability unnecessary for purposes	Same person is assigned **different pseudonyms**

In the temporal domain, the question poses itself over what time span data about the same person must be linkable. While in certain application fields, the purposes do indeed justify the need for very-long-term studies; in other fields, there exists some kind of “right to be forgotten” or something similar to a statute of limitation. An example for the former case may be long-term studies in medicine; an example of the latter case is a bad credit rating.

When designing temporal pseudonym domains, it is important to note that usually, it cannot be organized by snapshots. This is the case because snapshots usually have a significant overlap in their data content. Therefore, the linking between snapshots is possible even if different pseudonyms are used for each person. The linking can then be based on the data values instead. This does not only allow linking of data, but it also allows the linking of pseudonyms. This can then be used to link data that is contained only in a newer snapshot to data of an older snapshot.

The temporal compartmentalization thus has to be achieved indirectly by limiting pseudonym domains to cases, treatments, transactions, or similar. To avoid overlaps, their start or end date are then used to determine which temporal interval they belong to.

5.1.3.3.1 Technical considerations on compartmentalization

The following subsection discusses some technical aspects of how to reach compartmentalization though an adequate design of pseudonym domains.

It is assumed that in the domain of primary use, a stable long-term *person identifier* (**PID**) is used to fully identify persons. This PID represents the original basis for linking data belonging to the same person.

Since it may be required in certain cases to reverse the pseudonymization, in a given pseudonym domain (i.e., a compartment), there needs to be a relationship between the PID and the pseudonym.

This relationship can be expressed by a pseudonymization function that maps the PID of a person to its pseudonym.

Examples for pseudonymization functions include[65] [66]:

- A **lookup table** that allows to map in both directions between PID and a randomly chosen pseudonym. Such a lookup table obviously has to be **kept secret** in order to obtain compartmentalization.
- A (deterministic[67]) **encryption algorithms** together **with a secret key** such that the pseudonym is the encryption of the PID and reversely, the PID is the decryption of the pseudonym.
- A **keyed one-way-function** (such as an HMAC) where the pseudonym results from the application of the one-way-function to the PID. This function is mostly limited[68] to be used in the direction from the PID to the pseudonym.

Note that all these examples are based on a secret: the lookup table itself, the en-/decryption key, the key of the one-way-function. The secret represents the "additional information" mentioned in Article 4(5) GDPR. In particular, without this information, data subjects cannot be (re-) identified, i.e. their PID cannot be obtained from the pseudonym. Also, this additional information "is kept separately and is subject to technical and organisational measures" that protect the secrecy.

The presence of a secret is instrumental for successful pseudonymization (and thus compartmentalization). Without a secret, in Article 4(5) GDPR, there would be nothing to be kept separately and made subject to technical and organisational measures. Also, with a one-way-function that is not keyed (such as a hash or digest), the relation between PID and pseudonym can easily be established in one direction[69]. This can be used to obtain additional information about a person with a known PID from the secondary use domain. Further, if the possible values of the PID are relatively small[70], a person's PID can also be determined from the pseudonyms through a brute force attack (that may be aided with rainbow tables and similar techniques).

Secrets obviously become less secret the wider they are shared. This is relevant where linkability across data holders is necessary. The larger the groups of data holders that share a pseudonym domain and thus its secret, the weaker the compartmentalization or the "decoupling" of PID and pseudonym becomes.

This cannot be avoided where secret sharing is necessary to support necessary linkability. It increases the data protection risk, however. Where compartmentalization is weak, additional effort

[65] See also Konstantinos Limniotis, Hellenic Data Protection Authority, *Cryptography at the service of pseudonymization*, IPEN webinar 9 Dec. 2021:"Pseudonymous data: processing personal data while mitigating risks", https://www.edps.europa.eu/system/files/2021-12/03_konstantinos_limniotis_en.pdf (last visited 30/7.24).

[66] See also ENISA, Pseudonymisation techniques and best practices, Section 5, https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices, (last visited 30/7/24).

[67] Deterministic here means that the same input (clear text) always produces the same output (cypher text).

[68] Note that if the entropy of possible PIDs is small and the one-way-function is known, a lookup table (or at least rainbow tables) can be computed in a limited amount of time and thus enable the use of the function in the direction from pseudonym to PID.

[69] Note that this statement holds independently of the entropy of possible PIDs.

[70] Note that some possible PIDs may have a limited number of possible values (e.g., a consecutive number where the recently issued batch of numbers is known). Similarly, if the PID is based on a person's attributes such as name, gender, date and place of birth (see for example the Italian Tax Number), knowing the person often permits to drastically restrict the possible values of the PID. The Italian Tax Code can be calculated form a person's attributes online for example here: https://www.codicefiscaleonline.com/ (last visited 29/7/24).

and increased focus on other mitigation measures that limit the risk are necessary. In this sense, linkability "comes at a cost"; where linkability is unnecessary for the purposes, it is much cheaper overall to start with a strong compartmentalization of the data space in step 1, i.e., during the provisioning of secondary use data.

5.1.3.4 Other Considerations

The following discusses a data protection specific consideration beyond the principles described in Section 2. Namely it addresses the legal basis that permits the secondary use of data. It does so in the context of architectural option (iii) of snapshots.

In some data spaces or cases, the secondary use of certain data may require consent of data subjects. In this case, **consent can be withdrawn** by the data subject at any point in time (see Article 7(3) GDPR).

Similarly, some data spaces foresee that data subjects can **opt-out** of secondary use by data subjects[71].

Withdrawal of consent and opt-out are very similar. It can likely be assumed that Article 7(3) applies to both cases, stating that "[t]he withdrawal of consent **shall not affect** the lawfulness of **processing** based on consent **before its withdrawal**".

It is up to legal analysis how to apply this to the snapshot-based provisioning of secondary use data. There are two possibilities:

(i) A withdrawal of consent or an opt-out affect past snapshots and all still personal instances of data that were derived from past snapshots[72], or

(ii) a withdrawal of consent or an opt-out affect only future snapshots, while leaving past snapshots and its derive instances unaffected.

Technically and organizationally, the first option is likely complex and costly[73]. Also, the first option is incompatible with a possible reproducibility requirement (see Section5.1.1).

[71] See for example, the European Parliament wanting an opt-out possibility in the EHDS, https://www.europarl.europa.eu/news/en/press-room/20231208IPR15783/ep-supports-creating-eu-health-data-space-to-boost-access-to-data-and-research (last visited 26/7/24), in the section "Stronger safeguards for sensitive data".

[72] Such derived instances could have been transferred outside the data source, for example to a health data access body in the EHDS.

[73] The cost here is expected to lie mostly in the propagation of data deletion requests across a data space to all possible intermediaries (such as health data access bodies) and data users whose data instances contain still personal (i.e., not anonymized) data of the data subject).

5.2 Step 2: Provisioning the required data for a specific data use

This section discusses data protection of the processing aspect concerned with provisioning data for specific data use for a single data use request (called data access application in the EHDS proposal). As was shown in Figure 4, this step involves data holders and optionally one to multiple intermediaries (in the EHDS, these intermediaries are *data access bodies*). The discussion focuses on selected measures for different families of use cases. As shown in Table 6, the families of use cases are defined by the partition of the required data and the level of identification that is required.

Table 6: Families of use cases and selected measures.

Data space partition	Level of identification	Suitable Measure
Horizontal only	Individual-level data	*Section 5.2.1 and 5.2.2:* • **Data goes to analysis** • 2nd-level pseudonymization for compartmentalization. • Caching with garbage collection
	Aggregated data only	*Section 5.2.3:* • **Analysis goes to data** (*federated analysis / distributed learning*)
Vertical	Any data	*Section 5.2.4:* • **Data goes to integration point** (thereafter same as horizontal integration) • One-time-join-key for compartmentalization

5.2.1 Second-level pseudonymization for compartmentalization

This section addresses the use case where data users can show that they need pseudonymous individual-level data. This can for example be the case for explorative analysis where only looking at concrete individual cases, the necessary insights can be gained.

In this case, the only possibility is to move the data to the analysis (i.e., the data user); the option of moving the analysis to the data is infeasible.

The present subsection assumes that data have to be integrated only horizontally. How to handle vertical integration will be discussed in Section 5.2.4.

When only horizontal integration is necessary, two measures can be used in support of data protection principles. In particular,

(i) second-level pseudonymization can be used to support compartmentalization in support of the principle of purpose limitation and

(ii) the data can be cached with a garbage collection mechanism that enforces timely deletion in support of the principle of storage limitation.

The remainder of this subsections discusses these measures in detail.

5.2.1.1 Setting

This subsection provides more detail about the setting in which these measures are implemented.

Once data holders have completed their initial provisioning of data for secondary use, in the case where data user requests individual-level pseudonymous data, this data has to be adequately put at their disposition. The following discussion looks at this step. It starts from the initial snapshots that were provisioned by data holders for the entire family of all supported purposes. The discussed step now only needs to support the purposes pursued by a single requesting data user. Figure 6 illustrates the scope of the discussion.

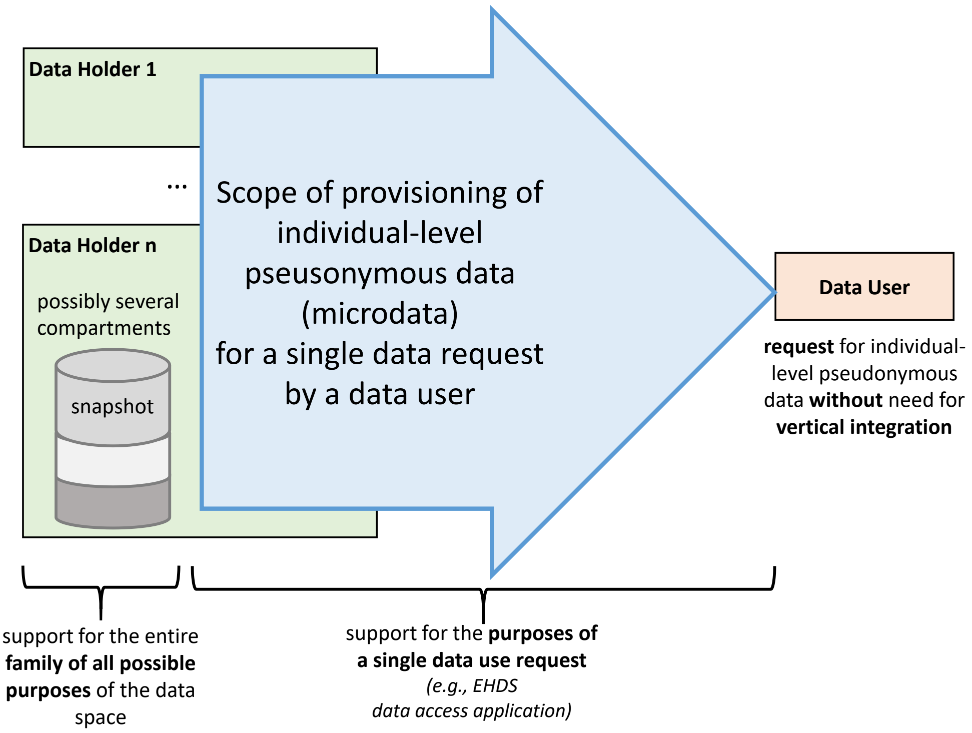

Data Holder 1
…
Data Holder n
possibly several compartments
snapshot
Scope of provisioning of individual-level pseusonymous data (microdata) for a single data request by a data user
Data User
request for individual-level pseudonymous data **without** need for **vertical integration**
support for the entire **family of all possible purposes** of the data space
support for the **purposes of a single data use request** *(e.g., EHDS data access application)*

Figure 6: The scope of the provisioning for a single data request starts after the provisioning of snapshots.

Note that several data holders can be involved. Also, the relationship between data user and data holders can be direct or make use of one or several intermediaries. Figure 7 shows a direct relationship; Figure 8 a topology with a single intermediary; and Figure 9 one with two levels of intermediation. Note that the levels of intermediation are not limited in a data space. For example, in a federally organized member state, there may well be the levels of province, nation, and EU, i.e., three levels.

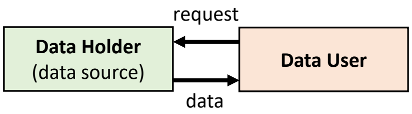


Figure 7: Data user issues request for data directly to data holder.

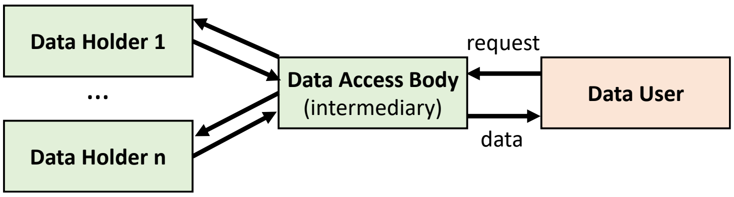


Figure 8: Data user issues request for data via a single intermediary.

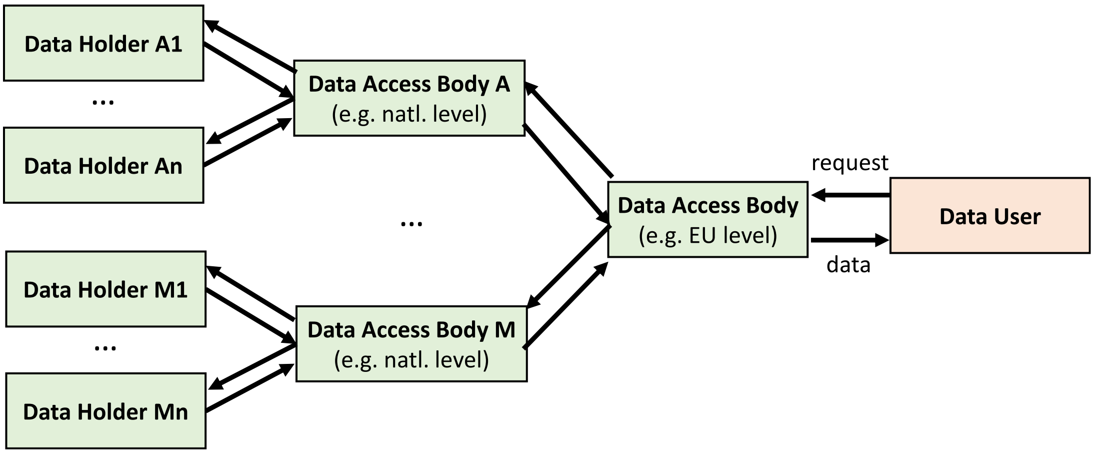


Figure 9: Data user issues request for data via a chain of intermediaries.

The data protection principles are all concerned with minimizing or limiting necessary processing activities[74] to what is actually necessary for each actor's purposes. Applied to the multiple actors of an intermediation topology, it is evident that intermediaries should handle only the data that are actually necessary to process a request by a data user.

Considering the three principles of *data minimization*, *purpose limitation* and *storage limitation*, it is clear that the aspects of data that are minimized to what is necessary are determined by the following aspects:

- Information content (not more than the data requested);
- identifiability (including need for pseudonymization reversal, individual-level vs. aggregate pseudonymous data); and
- linkability of pseudonyms (i.e., the minimally sized compartments or the maximal level of compartmentalization).

The necessary level of identification is mostly determined by the data user's request. In this section, the request requires individual-level pseudonymous data. Both, the case with and without the need for pseudonymization reversal (i.e., re-identification of data subjects based on their pseudonyms) is considered.

The minimization/limitation of data "capabilities" along a chain of actors is shown in Figure 9. Here, the data capacity of snapshots at data holders represent the base line. The reduction of data

[74] Here processing activity is used in a wider sense, encompassing data, the actual processing (in a narrower sense of functionality), and the time of storage or keeping the data in a form that permits identification.

capacities along the line limits the overall risk (e.g., by attacks and breaches) and takes into account that in contrast to intermediaries, data users are not fully trusted[75] actors in the data space.

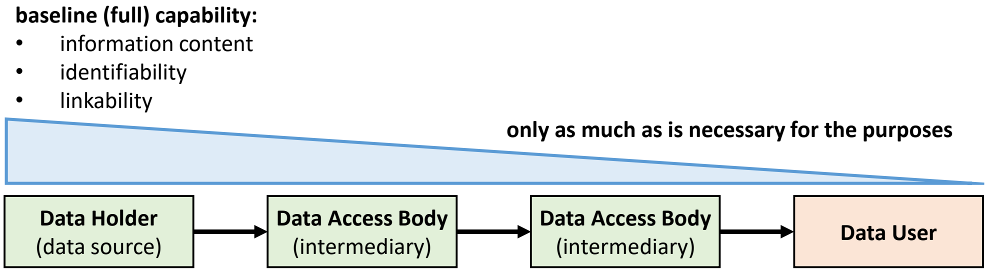

baseline (full) capability:
information content
identifiability
linkability
only as much as is necessary for the purposes
Data Holder (data source)
Data Access Body (intermediary)
Data Access Body (intermediary)
Data User

Figure 10: Data protection requests the minimization of data "capabilities" along the chain of actors involved in serving a user request.

In support of the discussion, it is useful to look in more detail at the anatomy of a data use request[76]. As discussed in Section 5.1.2, atomically addressable data elements at an individual data holder are identified the triple formed of the following components:

- DataHolderID
- TimeSpec (to identify the snapshot)
- CatalogID (to identify the category of information)

The vocabulary used for this identifier is standardized and coordinated in a data space. Thus, all data holders consistently use compatible schemes to specify the time of snapshots and use the vocabulary of a data-space-level catalog to identify categories of information.

In this setting, a request of data for a specific data use logically includes the following elements:

- A (potentially very large) list of all atomic data element (identified by the triple postulated above)
- Information about which pairs of atomic data elements need to vertically integrated (i.e., linked on pseudonyms from a single domain).

In the context of this section, the latter point refrains from requiring any vertical integration.

This leaves a potentially very large enumeration of atomic data elements. While such an enumeration is logically valid, practically for a data user, such enumerations are cumbersome and difficult. Therefore, the data space must foresee “named groups” of components. Of particular interest are named groups of data holder and possibly also named groups of several information categories that are frequently used together.

In the case of intermediaries, the present discussion assumes that a given data holder solely interacts with a single intermediary; a generalization to multiple intermediaries is then straight forward. The scenario with a single intermediary is for example the case for schemes whose topology follows a single level of administrative (and thus geographical) subdivision. Here, relations between data

[75] Note that a data space can vet the legitimacy of data users and thus extend a certain degree of trust.
[76] In the EHDS, a data use request is called “data access application” and is described in its Article 45.

holder and intermediary could for example be defined by a subdivision of a country provinces. Similar, the province-level intermediaries could then interact solely with a single Member-State-level intermediary. All Member-State intermediaries would then interact with a single intermediary at European level.

Such a hierarchical topology implies groups of actors such as all data holders interacting with a given province-level intermediary. Such groups can further be “filtered” into smaller subsets by using criteria on the attributes of data holders. A prime example would restrict the *type* of data holder.

By providing named groups, data users’ requests can be largely simplified and shortened. A data holder specification in a data use request can then hierarchically “decomposed”, traversing the hierarchy of intermediaries, to yield a (potentially very large) set of individual DataHolderIDs.

Since a data space may cater to a number of parallel or concurrent data use request from different data users, the relationship between multiple such requests is relevant. The relevance is for example important when considering storage periods, caching, and guaranteed deletion after the purposes are fulfilled. It can also help when considering ways to foster efficiency and the avoidance of duplications. From a data protection point of view, the latter can reduce the “attack surface” or “exposure time”.

5.2.1.2 Technical and Organizational Measures

For the scenario outlined above, this subsection discusses two measures in support of data protection principles, namely (i) compartmentalization through 2nd-level pseudonymization in support of purpose limitation and (ii) caching with garbage collection in support of storage limitation.

5.2.1.2.1 Compartmentalization through 2nd-level pseudonymization

Snapshots were created in support of a broad family of purposes possible in the given data space; in contrast, the task at hand caters solely to the purposes of a single data use request. Hence, the data to provision in this step shall not serve broad families of purposes, but must cater only to the purposes of a single use request. Data for distinct use requests shall be separated such that the data for one request cannot be easily used for other purposes of other requests.

As illustrated earlier, this requires that data serving one set of purposes cannot be easily linked to data related other purposes. This is achieved by using compartmentalization where different compartments use different pseudonym schemes (i.e., pseudonym domains).

In particular, considering the assumed absence of vertical integration, different data holder have no need to link data belonging to the same person. Similarly, different data users have no business of linking “their” data together.

The mechanism to create a distinct compartment (and thus pseudonym scheme) is 2nd-level pseudonymization. Here, in the same way as the initial pseudonym of the snapshot was derived from the long-term person identifier (PID)[77], a secondary pseudonym is derived from the primary pseudonym.

[77] See Section 5.1.3.3.1 that is part of Step 1 “provisioning for secondary use”.

As was the case for the primary creation of pseudonyms, where pseudonymization reversal is required, the secret used in the 2nd-level pseudonymization must also be kept[78].

The best way of organizing 2nd-level pseudonymization is that every actor employs it systematically prior to transferring pseudonymous data to an external actor. This then breaks the "backlink" to the (usually larger pool of) data managed by the actor. Such a backlink can then only be re-established by explicit and controlled pseudonymization reversal involving all actors along the chain. This approach is illustrated in Figure 11.

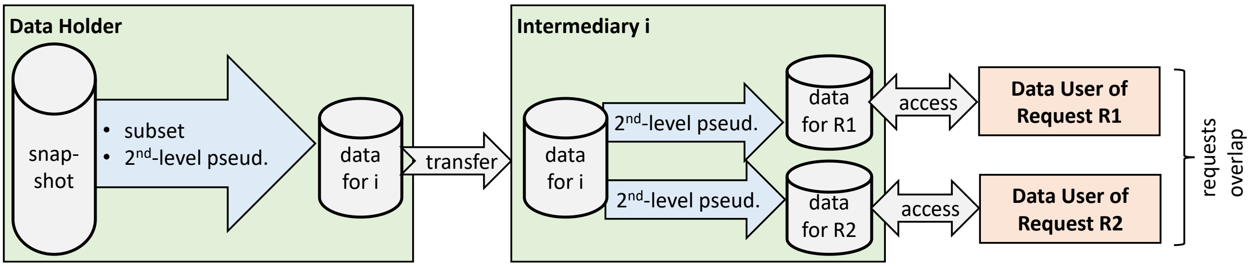


Figure 11: 2nd-level pseudonymization before any transfer to an external party.

The figure illustrates how this scheme of 2nd-level pseudonymization permits the use of a single data set (i.e., "data for i") for multiple purposes towards the right. The data for different purposes (uses) toward the right can still be separate before further transfer into distinct compartments (that use a distinct pseudonymization scheme). This property may be important for scalability and efficiency.

From a security and data protection point of view[79], it reduces the number of necessary transfers and the storage of redundant copies. By doing so, it reduces the attack surface and exposure time and thus the likelihood of breaches.

Since this scheme creates the final pseudonyms in multiple steps (prior to every transfer to another party), also the optional pseudonymization reversal must be executed in multiple steps. This is facilitated by the fact that the chain of re-identification follows that of requesting re-identification.

To enable pseudonymization reversal, every party that executes a primary or secondary pseudonymization must keep the corresponding pseudonymization secret. This can be achieved by including the pseudonymization secret in the metadata of processing a use request[80]. Use-specific pseudonymization secrets are typically random.

This scheme of multi-step pseudonymization enables the optional pseudonym reversal for two distinct scenarios:

- The data subject need to be notified about results of the secondary use of its data. In this case, the re-identification chain ends at the data holder who finally re-identifies the fully identified person ID (PID) from the initial pseudonym used in the snapshot.
- Exploratory data analysis by data users shows that additional (pseudonymous) data about a given data subject is required to proceed in the analysis or yield conclusive results. In this

[78] Separately and protected by suitable technical and organizational measures.

[79] Note that data protection requires security. See Articles 5(1)(f) and 32 GDPR.

[80] Note that in this approach, the per-use pseudonymization secrets should be stored in an encrypted form that prevents unauthorized pseudonym reversal.

case, the re-identification of the PID is unnecessary. Instead, the re-identification chain can end in the snapshot where additional data is available under the same pseudonym scheme.

From a technical point of view, the proposed multi-step pseudonymization can be fully automated behind a request interface. Every request for individual-level pseudonymous data without a need for vertical integration across multiple data holders then automatically executes the following steps:

- The data with an initial pseudonym scheme is extracted and temporarily stored.
- A random key (i.e., transfer-specific pseudonymization secret) is generated.
- If pseudonymization reversal is required, that key is encrypted and kept as part of the meta data of the request.
- The extracted data are converted by an on-the-fly 2nd-level pseudonymization based on the generated key. This operation is expected to be relatively inexpensive.
- The originally extracted data is deleted from temporary storage.
- The 2nd-level pseudonymized data is transferred in the response to the request.

Figure 12 summarizes the measure from the point of view of a single intermediary and integrates it with the task of horizontal integration. The role of the intermediary then consist of the following tasks:

- Decomposition of data requests towards the data source (left in the figure),
- horizontal integration of the data obtained from different sources, and
- 2nd-level pseudonymization of the data towards the data use (right in the figure).

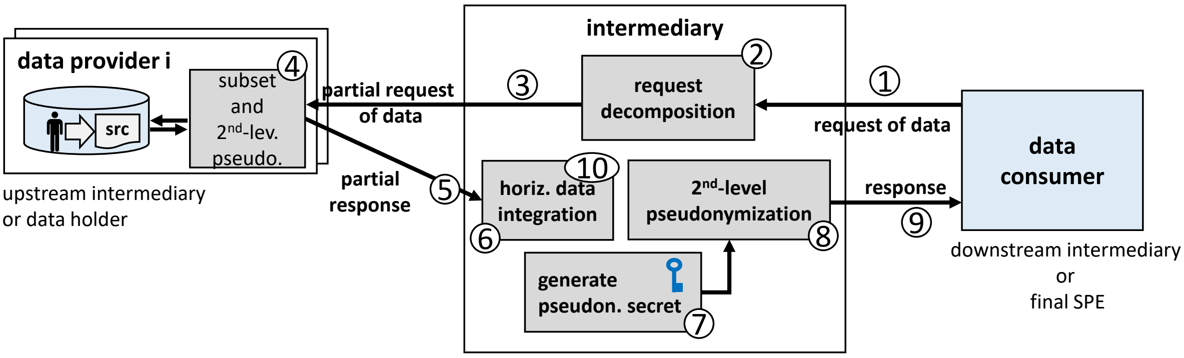


Figure 12: Role of intermediaries in individual-level data request with horizontal integration.

In particular, the figure shows the interaction of an intermediary with a downstream data provider (i.e., a "lower-level" intermediary or a data holder) on the left and an upstream data consumer (i.e., a "higher-level" intermediary or a data user within a Secure Processing Environment (SPE)) to the right.

Here, the following processing steps take place:

(1) The data consumer issues a request for data using group names for data providers.
(2) The intermediary decomposes this request into multiple partial requests that can be satisfied by single data providers implied in the group names.
(3) This partial request is sent to the competent data provider.

(4) The data provider extracts the necessary subset of data (and performs a 2nd-level pseudonymization specific for this request).
(5) The data provider sends this 2nd-level pseudonymized subset to the intermediary.
(6) On receiving all constituent subsets, the intermediary (horizontally) integrates these into a single data set.
(7) The intermediary then generates a random key suited as a pseudonymization secret.
(8) The intermediary uses this key for a 2nd-level pseudonymization (that is specific to the data consumer) of the integrated data set.
(9) The resulting data set is transferred as a response to the data consumer.
(10) The integrated data are no longer required and can be deleted[81].

Figure 12 can be modified for the case, where the party receiving the request of data has this data themselves locally. These parties include data holders who manage data in pseudonymized snapshots and intermediaries who find the requested data in their cache (see next section). The modification is shown in Figure 13.

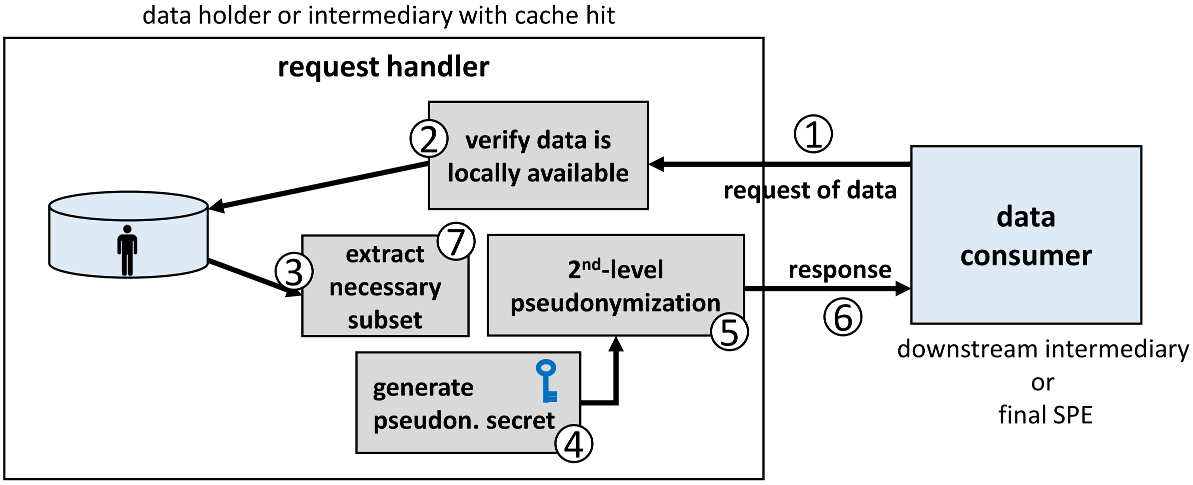

data holder or intermediary with cache hit
request handler
② verify data is locally available
⑦ extract
③ necessary subset
2nd-level pseudonymization ⑤
generate pseudon. secret ④
①
request of data
response
⑥
data consumer
downstream intermediary
or
final SPE

Figure 13: Request handler with local data.

In detail, the figure shows the following flow:

(1) The data consumer sends a request for data to the request handler, i.e., an intermediary or a data holder.
(2) The request handler verifies that the requested data is available locally.
(3) If this is the case, the necessary subset is extracted.
(4) A pseudonymization secret specific for this given request is generated.
(5) This secret is then used for a second-level pseudonymization of the extracted subset.
(6) The result is then transferred as response to the data consumer.
(7) The extracted subset of data is now deleted.

[81] Note that deletion here can be replaced by "more intelligent" caching with timely deletion guaranteed through an integrated "garbage collector". This is described in the following section.

The above two figures show the cases where a request handler is either in possession of none or all of the requested data. Evidently, there are also mixed cases, particularly for intermediaries who find only a subset of the necessary data in their cache (i.e., it is a partial cache hit). This situation is not shown in the figure but is straight forward to extrapolate from the concepts. In this case, also a decomposition of the data specification may be necessary. More details on caching are provided in the next subsection.

5.2.2 Caching with garbage collection

This subsection discusses a second technical measure in support of storage limitation. Namely, it proposes an analogy to caching with garbage collection. It is applicable to any use cases where data has to be transferred to the data use (by a requesting data user) and requires a redundant (temporary) storage by an intermediary[82].

Storage limitation states that personal data shall be stored only as long as necessary for the purposes. In the context of requesting pseudonymous data through several intermediaries, several redundant copies of the data may be created. A systematic management of such copies in order to guarantee timely deletion is therefore necessary.

Depending on the data space and its usage, it can be common that requests by data users have a significant overlap in the necessary pseudonymous data. It is therefore intuitive that where possible, the common data of requests are transferred only once in support of multiple overlapping requests.

A total absence of coordination between request would lead to additional cost (and energy consumption) and longer response times. From a security (and data protection) point of view, it would increase the attack surface, exposure time, and thus risk. Such an approach is therefore not optimal.

On the other extreme, a preventive collection and long-term storage of data by intermediaries would likely be disproportional to the actual need of coordination. It would also be an approach that is difficult to justify from a data protection and also political point of view. Note for example that the EHDS in its Article 46(4) foresees that intermediaries request data from data holders only once an approved data access application is in place. In other words, this contradicts an approach of preventive collection by intermediaries.

The optimum solution must therefore find a balance between these two extreme approaches, namely between a total lack of coordination of overlapping data use requests and a preventive coordination and collection.

To understand the problem, it is important to note that overlapping requests for data usually occur at different points in time. When a first request comes in, it is therefore not yet clear whether and when overlapping requests will be issued. Depending on the transfer and processing cost that were spent to obtain a temporary data set at an intermediary, it is reasonable to wait a reasonable period of time[83] after satisfying the first request before deleting the data. From a data protection point of a reasonable extension of the minimally necessary storage time can potentially eliminate the risks of

[82] This intermediary is for example the final data access body in the EHDS who provided access to this pseudonymous data to data users within a secure processing environment. It could however also be other intermediaries who require temporary storage of redundant copies of data for example for the purpose of vertical integration.

[83] What is a reasonable period of time depends on the characteristics of the data space and its (current) usage and its determination may require legal balancing.

repeated data transfer and avoidable additional redundant storage of temporary and intermediate data sets.

The desired compromise solution evidently has the characteristics of a cache:

- The temporally extended storage is time-limited and not guaranteed,
- In the period in which data are cached, one or potentially a multitude of consecutive requests for the same data can be executed more efficiently.

Section 5.2.1.1 has introduced the concept of “atomically addressable data elements”[84]. If the cache stores these elements, the overlap in data use requests can always be expressed as a set thereof.

Data protection requires that such cached data should be deleted[85] once it is no longer necessary for the purposes. In this context, this means as long as they are no longer required for at least one data users. Evidently, there is an analogy between the number data user who require the data and a reference counts in a garbage collector.

Where intermediaries implement a cache of pseudonymous data, they should thus also implement the equivalent of a garbage collector. This constitutes a second line of defense compared to relying on triggering deletion when a data use is completed. Should the latter be forgotten or fail for some reason, the garbage collector then still enforces a global policy that ensures timely deletion of cached data across all data users.

In summary, where a data space needs to support the use of individual-level pseudonymous data, the only option is to bring the data to the user. This implies the storage of redundant copies of subsets of data by intermediaries. Such storage has to be managed systematically in order to guarantee efficiency and timely deletion of no longer needed data. A promising systematic approach uses the concepts of caching and garbage collection.

5.2.3 Federated Analysis

The above scenarios described cases where the data is brought to the analysis. This is unavoidable if data users directly need accesses to individual-level data. In scenarios, where data users only need aggregated data, there is the possibility to bring the analysis to the data. This is called federated analysis. From a data protection point of view it is far superior to bringing data to the analysis.

Providing access only to aggregated data is different from providing anonymous data. In particular, in contrast to using anonymous data, the aggregating analysis is conducted on a complete data set (without suppression of outliers) and truthful data (without the addition of random noise as for example in differential privacy approaches).

Figure 14 illustrates the concept of federated analysis. In addition, it description uses the simple example of an analysis of the average height of data subjects.

[84] These are DataSetIDs where all group names are decomposed.

[85] At least after a justified delay. Note that transforming them into a form that no longer permits identification is not applicable in this context.

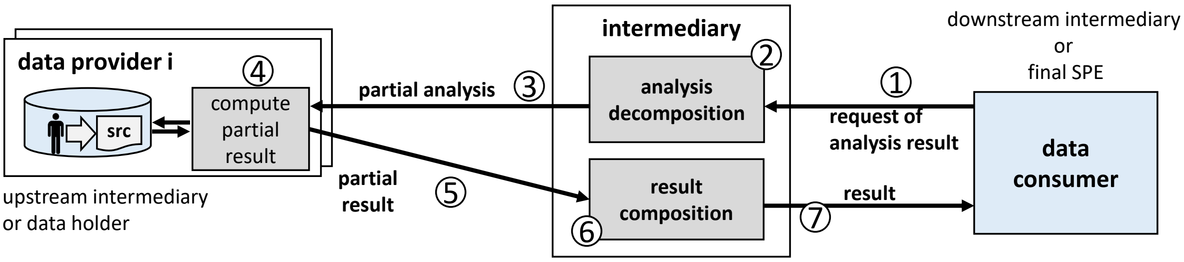


Figure 14: Federated Analysis.

The flow of exchanges in this figure is described in the following:

(1) A data consumer requests an aggregate analysis from an intermediary. In the illustrative example, the intermediary may be national and the request is for the average height of all data subjects in the nation.

(2) The intermediary decomposes the request into a multitude of partial requests that can be answered by a single constituent data provider. In the example, a partial request would ask a single data provider the average height and number of data subjects.

(3) Each partial request is sent to the competent data provider.

(4) The data provider then applies the requested partial analysis to its data set. In the example, the data provider computes the average height and the number of data subjects.

(5) This partial analysis result is sent as response to the intermediary.

(6) The intermediary now collects all partial results and composes the overall analysis result from it. In the example, this constitutes taking the weighted average of partial average heights with the number of data subjects as weight. Obviously, the partial results are no longer needed and can be deleted.

(7) This result is then sent as response to the data consumer.

Federated analysis has seen large scale use at least in the areas of statistics and machine learning. The reported large-scale operational deployments imply a high level of technology readiness.

In statistics, federated analysis on top of the ubiquitous open source package *R*[86] has is in wide use particularly in biomedicine. The resulting software suite is called *DataShield*[87] and is developed in an open source model itself. It implements the basic R commands to add decomposition and decomposition to the functionality. Its implementation also enforces a minimum level of aggregation that can be configured. Due to this approach, DataShield is also compatible with R's graphical tools such as *RStudio*[88]. Practical use cases of DataShield, particularly in biomedicine and life sciences are for example reported at the yearly *DataSHIELD Conference*[89]. Also the German

[86] https://www.r-project.org/ (last visited 4/12/24).
[87] https://datashield.org/ (last visited 4/12/24).
[88] https://posit.co/products/open-source/rstudio/ (last visited 4/12/24).
[89] https://www.mathematics-and-life-sciences.uni-bonn.de/en/datashield-conference-2024/datashield-2024-programm (last visited 4/12/24)

National Research Data Infrastructure for Personal Health Data (NFDI4Health)[90] uses DataShield at a large scale[91].

DataShield also provides evidence that aggregate-level data is not anonymous but rather pseudonymous and requires the necessary "access governance"[92]. For example, an identification attack[93] that is also applicable to DataShield is acknowledged[94] by the DataShield community who emphasizes that DataShield requires "access governance"[95].

Federated learning[96] is an established field of machine learning. It is already in operational use at large scale[97]. For example, it has been used to create model for self-driving cars[98], for on-device intelligence[99], in health[100], and in robotics[101].

5.2.4 Vertical (or mixed) data integration

The discussion above focused on scenarios where data is only horizontally partitioned and no vertical integration is required. This section extends the above measures to include vertical integration of data.

A simple example is used to illustrate the situation. Here, two types of approved analysis requests are considered:

(i) The analysis requires actual individual-level body mass index[102] (BMI) values for pseudonymous data subjects, or

(ii) the analysis requires the average (or a histogram of) of the population's BMI (i.e., an aggregate pseudonymous value).

It is further assumed that the data space constitutes a mixed partition (i.e., the most general case). The assumed data space is visualized in Figure 15. Here, a geographic (for example regional)

[90] https://www.nfdi4health.de/en/ (last visited 4/12/24).
[91] https://www.nfdi4health.de/en/service/fostering-collaborative-research-environments-using-datashield.html (last visited 4/12/24).
[92] Note that in the EHDS proposal, such "access governance" is implemented for example by the necessity of approval of data access application (see Art. 46(3) EHDS proposal) and by a secure processing environment (see Art. 50 EHDS proposal).
[93] Huth, Manuel & Gusinow, Roy & Contento, Lorenzo & Tacconelli, Evelina & Hasenauer, Jan. (2022). Accessibility of covariance information creates vulnerability in Federated Learning frameworks. 10.1101/2022.10.09.511497, https://pubmed.ncbi.nlm.nih.gov/37647639/ (last visited 4/12/24).
[94] https://datashield.discourse.group/t/vulnerability-in-federated-analysis-software/622 (last visited 4/12/24).
[95] https://datashield.discourse.group/t/statement-datashield-disclosure-controls-and-mitigation/628 (last visited 4/12/24).
[96] https://en.wikipedia.org/wiki/Federated_learning (last visited 4/12/24).
[97] See for example https://en.wikipedia.org/wiki/Federated_learning#Use_cases (last visited 4/12/24).
[98] *Elbir, Ahmet M. and Sinem Coleri. "Federated Learning for Vehicular Networks." ArXiv abs/2006.01412 (2020),* DOI:10.48550/arXiv.2007.13518.
[99] Konečný, Jakub & McMahan, H. & Ramage, Daniel & Richtárik, Peter. (2016). Federated Optimization: Distributed Machine Learning for On-Device Intelligence. 10.48550/arXiv.1610.02527, https://arxiv.org/abs/1610.02527 (last visited 4/12/24).
[100] Karargyris, A., Umeton, R., Sheller, M.J. *et al.* Federated benchmarking of medical artificial intelligence with MedPerf. *Nat Mach Intell* **5**, 799–810 (2023). https://doi.org/10.1038/s42256-023-00652-2.
[101] B. Liu, L. Wang and M. Liu, "Lifelong Federated Reinforcement Learning: A Learning Architecture for Navigation in Cloud Robotic Systems," 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 2019, pp. 1688-1695, doi: 10.1109/IROS40897.2019.8967908.
[102] The BMI is defined as the ratio of the weight and the square of the height.

subdivision into administrative units (e.g., regions) is depicted. An intermediary at a higher (e.g., national) level connects to multiple such units. Each unit contains multiple types of data holders that are represented by disks. The disks marked with “h” collect height data for all inhabitants of the unit; those with “w” collect weight data. Evidently, in order to calculate a BMI of a person, data from two distinct data holders are required. Also, this data must be vertically integrated since the weight and height that are put into relation need to belong to the same person.

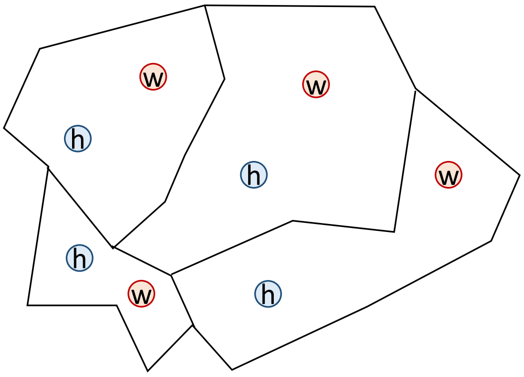


Figure 15: Illustrative scenario for a mixed data partition.

In the measures for horizontal integration described in the previous subsections, every data holder used second-level pseudonymization of disclosed data in order to lock them into a single compartment and prevent the linking of pseudonyms across compartments from different data holder. When vertical integration is required, this compartmentalization does no longer work. Instead, in the above scenario, data holders of the same administrative unit need to use the same pseudonym for each given person, i.e., they need to use a shared compartment.

Linking pseudonyms across administrative units is unnecessary, in contrast. The principle of purpose limitation thus mandates that data holders of different administrative units use distinct compartments that prevent the linking of pseudonyms.

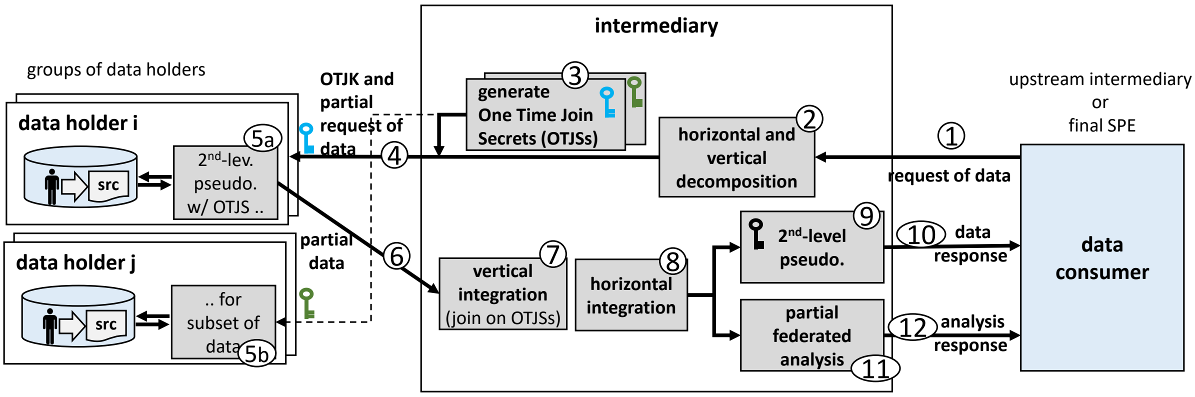

intermediary
groups of data holders
data holder i
src
5a
2nd-lev. pseudo. w/ OTJS ..
OTJK and partial request of data
4
3
generate One Time Join Secrets (OTJSs)
2
horizontal and vertical decomposition
1
request of data
upstream intermediary or final SPE
data holder j
src
.. for subset of data
5b
partial data
6
7
vertical integration (join on OTJSs)
8
horizontal integration
9
2nd-level pseudo.
10
data response
partial federated analysis
11
12
analysis response
data consumer

Figure 16: Vertical integration with a One Time Join Key (OTJK).

Figure 16 illustrates how this can be achieved. In particular, the processing steps are the following:

(1) A data consumer requests either (i) individual-level pseudonymous data or (ii) aggregate-level pseudonymous data. In the example, this would correspond to (i) pseudonymous weight and height data and (ii) an average BMI data.

(2) The intermediary decomposes the request. In the example, this results in two partial requests per administrative unit—one partial request for height data and one for weight data.

(3) For every domain of vertical integration, i.e., administrative unit in the example, the intermediary then generates a random *one time join secret* (OTJS).

(4) It then sends these OTJSs paired with the partial request for data to the competent data holders.

(5) The data holders then extract the necessary subset of data and perform a second-level pseudonymization with the received OTJS. In the example, the data holders of the same administrative unite hold different kinds of data (i.e., weight or height, respectively) but use the same OTJS. Consequently, both generate the same pseudonym for the same person. (5a) and (5b) represent different administrative units which use different OTJSs.

(6) The resulting compartmentalized data is then sent by the data holders to the intermediary.

(7) The intermediary collects the partial data of all data providers who share the same OTJS and integrate the data vertically by joining[103] the data on the OTJS. In more detail, in the example, the weight value and the height value belonging to the same pseudonym are grouped together.

(8) After the vertical integration, the joined data can be horizontally integrated. In the example, this brings all joint height and weight data together across all administrative units.

(9) In the case (i) where the data consumer requested individual-level data, the integrated data has to undergo another 2nd-level pseudonymization to create a specific compartment for the request (and thus data user) at hand.

(10) After this 2nd-level pseudonymization, the data can be sent to the data consumer.

(11) In the case (ii) where the data consumer requested aggregate-level pseudonymous data, the intermediary has to compute either the requested result (such as an average BMI in the example) or a partial response that is part of a federated analysis.

The figure shows the point of vertical integration and the point of partial federated analysis both at the shown intermediary. In more complex (composed) cases where there are chains of intermediaries, both these points must be located as close as possible to the data sources (i.e., data holders). This maximized compartmentalization and thus the principle of purpose limitation. From a data protection point of view it minimizes risk.

[103] In DBMS, this corresponds to an inner join on the pseudonym created based on a shared OTJS as common key.

5.3 Step 3: Providing access to pseudonymous data

In step 2 that was discussed in the previous section, pseudonymous data was provisioned for a specific data use (based on a request/application by a data user). This data has now arrived in an integrated and compartmentalized state at the final intermediary. This intermediary directly interacts with the requesting data user, providing either **on premise or remote access** to the provisioned data.

The provisioned data is clearly pseudonymous. It therefore underlies[104] all the requirements of the GDPR. This includes for example *confidentiality*[105]. The main measure to fulfill these requirements is that of ***Secure Processing Environment*s** (**SPE**s). In particular, an SPE is provided by the final intermediary to control access by data users to pseudonymous data.

The remainder of this section discusses the measure of SPEs in more detail.

5.3.1 Secure Processing Environments

The following describes how SPE providers can control the actual processing by data users in order to enforce data protection compliance. In the discussion, the data protection principle of *purpose limitation* is in the focus.

The setting is described in the following. In response to a data use request[106], the required data are provisioned in a dedicated, use-specific compartment to the final intermediary. In the following, these data are called ***Use Specific Data*** (**USD**). Data users have declared in their data use request which purposes they pursue. This request has then been formally approved[107] by the competent body of the data space. Approval is a prerequisite for provisioning the required use specific data[108] to an SDE in the first place. The following is concerned with measures that limit a user's processing to what is actually necessary for these declared purposes.

Use specific data cannot simply be handed over to data users for processing (see Figure 17)[109]. In particular, in this case, it would be impossible to limit the processing to what is necessary for the declared and approved purposes. For example, it would be impossible to prevent the following examples of excessive processing:

- Assessing the market potential of a new pharmaceutical based on the prevalence of certain conditions in patients, instead of pursuing the declared purpose of academic medical research.

[104] Note that Recital 26 states this explicitly in its second sentence: "Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person."

[105] See the principle of "'integrity and confidentiality'" in Art. 5(1) (f) GDPR.

[106] For example, in the EHDS, such a data use request is called "data access application" and is the subject of its Article 45.

[107] For example, in the EHDS, the approval of data access applications is described in its Article 46.

[108] See for example Art. 46(4) EHDS proposal that states "Following the issuance of the data permit, the health data access body shall immediately request the electronic health data from the data holder."

[109] Note that Recital 54 EHDS states this explicitly: "Such secure processing environment should reduce the privacy risks related to such processing activities and **prevent the electronic health data from being transmitted directly to the data users**."

- Enriching the use specific data though linking[110] with external data to enable the following types of processing:
  - Types of processing that becomes possible thanks to the data enrichment but could not be supported by solely the provisioned use specific data; and
  - Processing with the purpose of re-identifying data subjects in the pseudonymous data[111].

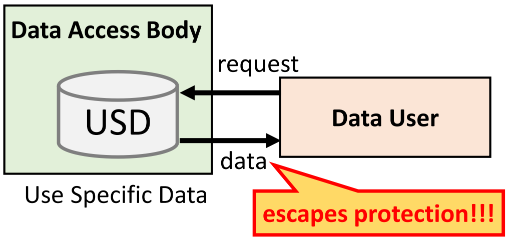


Figure 17: Simply passing use specific data to data users prevents any limitation of by data users.

To address these risks, the measure foreseen in the current legislation on data spaces are *Secure Processing Environments* (SPEs). In particular, they are defined in Article 2(20) DGA and take a prominent role in the EHDS (see for example its Article 50).

In its Article 2(20), the DGA defines Secure Processing Environment as follows: (Text within brackets, emphases, and footnotes added by author).

'[S]ecure processing environment' means the **physical** or **virtual environment** and organisational **means to ensure compliance** with Union law, such as Regulation (EU) 2016/679 [i.e. the GDPR], in particular with regard to data subjects' rights[112], intellectual property rights, and commercial and statistical confidentiality, integrity and accessibility, as well as with applicable national law, and to allow the **entity providing the secure processing environment** to **determine and supervise all data processing actions**, including the *display*, *storage*, *download* and *export* of data and the *calculation of derivative data* through computational algorithms;

Thus, an SPE can be implemented **physically or virtually**:

- In the former case, the SPE-provider dedicates premises equipped with the necessary IT infrastructure that can only be **accessed on premise**;

- in the latter case, the SPE is an interface through which the use specific data can be **accessed remotely**.

Physical SPEs obviously come at a significantly higher cost (to both, SPE-providers and data users) than virtual SPEs that permit remote access. Which implementation option to choose depends on the data protection risk represented by the use specific data. One factor that affects this risk is the level of identification of data subjects. In particular, individual-level pseudonymous data represents a higher risk than aggregate pseudonymous data.

[110] Note that the compartmentalization prevents the straight-forward linking on pseudonyms; it cannot prevent the linking on unique combinations of data values. Compartmentalization thus renders linking more difficult/costly and reduces the likelihood of certain matches, but it cannot prevent linking all together.
[111] Note that Article 44(3) EHDS prohibits data users from re-identifying data subjects.
[112] At least in the EHDS, the authors fail to see how SPEs are used in support of data subject rights listed in Chapter 3 GDPR, i.e., its Articles 13 through 23. In the authors' view, SPEs predominantly support the data protection principle of *purpose limitation* that is to a large part protected though confidentiality.

Figure 18 illustrates the use of an SPE and covers both of these options. Here, access to use specific data is only possible through the functionality and safeguards provided by the SPE. The functionality aspect can be understood by an analogy to smartcards that store secret keys: The keys are not directly accessible or extractable, but the provided functionality permits to use the key for example for encryption, decryption or signing. In this concept, the available functionality is provided and thus determined by the SPE-provider. Should data users be permitted to import their own functions[113] into an SPE, these must be subject to assessment and prior approval by the SPE-provider.

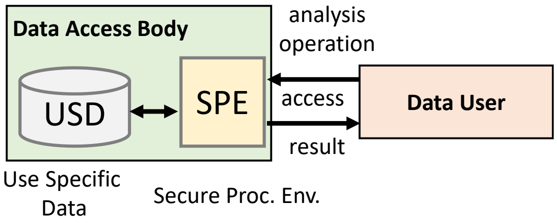


Figure 18: A secure processing environment can be used to control and supervise the processing operations executed by data users.

An SPE sits between data users and the actual data. Any access to the data requires functionality, for example in the form of commands or menu options. To perceive or use data, humans require functionality such as "view" (read), "edit" (read, write, modify), "copy" or "download", "visualize", or different forms of "analyze". This functionality is provided by the SPE. The SPE can thus control access to the data by restricting the functionality that is available to users.

Secure processing environments (SPEs) have to implement the following three requirements:

(1) **Prevent** pseudonymous **data from escaping the SPE** and thus prevent that they can be processed in absence of the measures that are implemented by the SPE.

(2) Since SPEs implement *pseudonymous processing*[114] that is defined in Art. 4(5) GDPR, they must implement "technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person". In other words, **SPEs shall prevent unauthorized**[115] **re-identification** of pseudonymous data. Prevention of (re-)identification includes the following aspects:

   a. Controlled access to the additional information that is intentionally kept to allow **pseudonymization reversal**. (See Section 5.2.1.2.1 and Art. 44(3) EHDS Proposal).
   b. Prevention of re-identification through **linkage with (external) additional information**, other than that subject to 2a.
   c. Prevention of re-identification through ***spontaneous recognition***[116] of data subjects by legitimate data users.

[113] This holds for both, functionality and auxiliary external data. An example for when the "import" of external functionality may be justified is the use of new, state-of-the art analysis methods (in the form of software) developed by a data user who conducts research. Such functionality cannot be provided by SPE-providers themselves but may be a legitimate and instrumental part of research and innovation.

[114] According to its definition in Art. 4(5) GDPR, *pseudonymization* is a manner of processing in which identification of data subjects is not possible except in controlled and intended ways.

[115] Note that SPEs are by all means compatible with intended re-identification that is, for example, foreseen in Art. 38(3) of the EHDS proposal.

[116] Spontaneous recognition is a type of identification by users who know a data subject such that their knowledge constitutes "additional information" that renders identification possible and involuntarily unavoidable.

d. Prevention of **reconstruction of individual-level data from aggregated data** since, for aggregate-level data, this is a necessary step towards re-identification.

(3) Measures in support of *purpose limitation* that **restrict the processing** by data **users to what is necessary for the purposes** that were declared in the data access request and subsequently formally approved by the provider of the SPE.

Measures suitable to implement these requirements are discussed in further detail in the following.

(1) Prevention of data from escaping the SPE

Measures to prevent pseudonymous data from leaving the SPE (and then be processed without its protection) are widely different for physical and virtual SPEs. In particular, physical SPEs permit much better control over the physical (i.e., HW) and logical means (i.e., SW and auxiliary data) that are at the disposition of data users. Further, physical SPEs permit a better supervision of the activities (i.e., executed functionality) performed by data users. In contrast, in virtual SPEs, only activities conducted on the SPE server are controllable; activities on the client-side cannot be practically restricted.

Table 7 gives an overview of measures to prevent data from escaping an SPE.

Table 7: Measures to prevent pseudonymous data from escaping SPE.

Measure	Phys. SPEs	Virt. SPEs	Effectiveness
Prohibit **HW for the capture, storage, or transfer of data**. For example, cameras, USB-sticks, and smartphones could be banned from entering/exiting the SPE. Printers must be avoided inside the SPE.	Yes	No	Preventive
Avoid **HW interfaces** necessary to capture, store, or transmit data. For example, USB ports are disabled.	Yes	No	Preventive
Prevent legitimate data users from **seeing data of other** legitimate data **users**. For example, work places in the SPE could be separated by room divider screens or monitors could by equipped with “anti spy” filters.	Yes	No	Preventive
Prevent **undesired network connections** to the user work stations of the SPE. For example, by firewalls, network intrusion detection, frequency jammers, air gapping.	Yes	No	Preventive
Prevent any **SW** (command or functionality) **suited to extract data**. For example, commands to copy, archive, upload, export, print, etc. data.	Yes	Servers only	Preventive

Supervision of activities by data users together with detection of anomalies and sanctioning of misbehavior. For example, video surveillance, oversight by staff, logging of electronic activities.	Yes	only interaction with servers	Less likely
Vetting and **authentication**[117] of data users.	Yes	Yes	Less likely
Codes of conduct or **binding obligations** for data users.	Yes	Yes	Less likely
Exit Gateway where any **data leaving the SPE** is checked and approved. This is for example necessary for the "results or output of the secondary use" that can leave the SPE solely if they "**only contain anonymised data**" (see Art. 46(11) EHDS Proposal).	Yes	only if created on server	Preventive

Physical SPEs evidently permit more and more effective measures. But they also come at a considerable cost to both, providers as well as data users. The decision to use a physical or a virtual SPE is thus likely informed by a risk analysis. In particular, highly sensitive data with a still substantial risk of re-identification likely require a physical SPE; in contrast, disclosure of moderately sensitive information at aggregate-level only may well be suitable to be managed in a virtual SPE.

Aggregate-level pseudonymous information can colloquially be seen as "almost anonymous". Aggregate-level information is for example disclosed when users can request only statistics over individual-level pseudonymous data. These statistics are then always computed over groups of multiple individuals.

Based on this consideration, it is likely that secondary use of individual-level pseudonymous data must be executed in a physical SPE; whereas virtual (remotely accessed) SPEs are suited for uses where only aggregate-level pseudonymous data is disclosed.

(2) Prevention of identification inside the SPE

The following discusses possible measures for preventing identification of data subjects during the processing of pseudonymous data inside of an SPE. This is a major characteristic required of pseudonymization, i.e., a manner of processing of pseudonymous data, in the GDPR (see Art. 4(5)).

The **main mechanism** to prevent identification is to **disclose only aggregate-level pseudonymous data** to data users. This was discussed in Section 5.2.3 above in the context of *federated analysis* and the *DataShield* system. Colloquially, aggregate-level information could be said to be "almost anonymous" since singling out of data subjects is no longer possible. Evidently, this drastically reduces the risk of re-identification compared to the case where individual-level data is disclosed and data elements pertaining to an individual are already singled out.

A more systematic discussion of measured is shown in Table 8. It addresses all possibilities of identification discussed in the introduction of the present section.

[117] See also Article 50(1)(a) EHDS proposal. Here, authentication is the basis for authorization of access (i.e., access control). Vetting can also be seen as a part of enrollment in an (physical or remote) identity management scheme.

Table 8: Measures to prevent the identification of pseudonymous data within an SPE.

	Type of re-identification	Measure	Effectiveness
Applicable to disclosure of **individual-level information**	Unauthorized pseudonymization reversal	**Well-defined conditions, procedures, and authorized personnel** for pseudonymization reversal	Preventive
	Linkage to external additional information	**Prevent uncontrolled data import to SPE** Note that secondary uses may legitimately require the import of auxiliary data. Prevention of uncontrolled import can be implemented through the limitation of HW interfaces, networks, or functionality.	Preventive
		Entry Gateway where any **data entering the SPE** is checked and approved. Imported data must not be suited for re-identification though linkage.	Preventive
	Spontaneous recognition	Restrict data **access to aggregate pseudonymous data only**	Preventive
		Preventively **assess likelihood of data users knowing data subjects** and prevent access in case of high likelihood.	Less likely
		Procedure to report occurrences of spontaneous recognition to SPE provider and take measures to control damage. For example, avoid that a data user learns additional information about a spontaneously recognized person.	Damage control
Applicable to the disclosure of **aggregate-level information**	Reconstruction (a necessary first step of the re-identification of aggregate pseudonymous information)	**Detect** collection of multiple data sets suitable for reconstruction **and prevent further access**. Denial of access comes in time to prevent the collection of a sufficient number of data sets. Detection can be per data user across multiple uses. Collusion of multiple data users is difficult to detect.	Preventive
		Detect suspicious collection of data sets **after the fact** and	Less likely

		investigate further. Uncovered cases of reconstruction trigger penalties.	

(3) Purpose limitation within SPEs

The following discusses measures to limit the processing of data users to the declared and approved purposes. This requires that this declaration of purposes is relatively detailed; very wide purposes as provided in Art. 34(1) of the EHDS proposal render the decision difficult whether a certain processing operation is indeed necessary. For example, what processing is necessary for "scientific research related to health or care sectors" (Art. 34(1)(e) EHDS proposal) it is difficult to decide whether a concrete processing step is indeed necessary.

According to Art. 45(2)(a) EHDS proposal, a data access application must contain "a detailed explanation of the intended use". It is reasonable to expect that this also includes a description of the pursued (anonymous) results and outputs that are subject of Art. 46(11) EHDS proposal.

Table 9 systematically lists possible measures to limit processing to what is necessary for the declared purposes.

Table 9: Measures of purpose limitation within an SPE.

Measure	Effectiveness
Preventive approval of an analysis expressed as an executable script. The most rigorous manner to limit processing to what is necessary is to have data users express their intended processing in the form of an executable script and execute it subject to prior approval. This is costly and incompatible with explorative analysis. It may require frequent amendments and additions.	Preventive
Logging of **processing activities** and checking them after the fact. Since excessive processing can be detected and sanctioned, this measure represents a deterrent for misbehavior.	Less likely
Restriction of available functionality (such as libraries)**.** For example, the R statistics language can load a wide variety of libraries (called packages, see https://cran.r-project.org/web/packages/available_packages_by_name.html). The libraries that are made available to data users can be limited to those necessary for their declared purposes.	Preventive of certain categories of excessive processing

The measures implemented by SPE providers discussed in the three subsections above can further be complemented by measures implemented by data users. Data access applications in the EHDS proposal explicitly contain "a description of the safeguards planned to protect the rights and interests of the data holder and of the natural persons concerned" (see Art. 45(2)(f)). Such safeguards can include:

- keeping client equipment secure,
- internal access restrictions by organizational data users for information obtained from a remote SPE, or
- measures to create awareness of adequate behavior among persons acting for an organizational data user.

The effectiveness of measures implemented by data users is always limited (from the point of view of SPE providers) since they fully depend on the trustworthiness of data users; malicious data users remain completely unaffected by such measures.

5.4 Step 4: Providing anonymized data to data users

The step described in this section is concerned with providing anonymized personal data for secondary use to data users. The following discussion takes a general approach to this problem. In contrast, the EHDS is limited to exclusively anonymous data in statistical format[118]. The discussion encompasses also the possibility of other kinds of formats (such as machine learning models or decision trees), but should be equally applicable to the EHDS.

In the EHDS, step 4 is typically related to a *data request* (see Art. 47 EHDS proposal). Here, data user request anonymous data. Since anonymous data is understood to pose no risk to the rights and freedoms of natural persons, it falls outside the scope of the GDPR. Data user can therefore directly obtain anonymized data in response to their request. The protection of an SPE is unnecessary in this situation. While not foreseen in the EHDS, being anonymous, this data could also be made public.

In the legal world, personal and anonymous are often understood at binary states. The EHDS acknowledges[119], however, that anonymity is usually not absolute but much rather, a residual risk of re-identification remains. This case is excluded from the discussion and partly subject of the forthcoming project Task 4.9.8. While excluded from the present discussion, the question how a data space can handle a possible "anonymization breach" is highly relevant. For example, what happens if a new kind of re-identification attack is discovered for an anonymization technique that has been used for an extended time in a data space?

The primary data protection objective of step 4 is that the anonymization is indeed effective. This can be supported by measures discussed in the sequel of this section.

For better understanding, it is important to highlight the differences between pseudonymous data as processed in an SPE and anonymous data. The main differences are summarized in Table 10. It also provides the main reasons for users to request one or the other kind of data.

Table 10: Comparison of pseudonymous and anonymous data.

	Pseudonymous data	**Anonymous data**
Truthfulness	truthful	Deviation from the truth due to noise injection or perturbation
Completeness	complete	Incomplete due to suppression of outstanding values
Level of detail	full detail	Reduced detail due to generalization

The remainder of this section lists possible measures in support of the effectiveness of anonymization.

[118] While this is somewhat unclear in Art. 47 and Recital (49) of the commission proposal of the EHDS, it is explicitly clarified in Art. 47(1) and Recital (49) in the compromise text at https://www.consilium.europa.eu/media/70909/st07553-en24.pdf (last visited 12/12/24).

[119] See Recital 64 EHDS proposal.

5.4.1 Full disclosure control

The creation of anonymous data in data spaces should be based on full disclosure control which is **more than just an "anonymization technique"**.

Disclosure control is best known from the field of statistics[120]. It goes beyond just an "anonymization technique" in the following two ways:

- **Aggregation** of individual-level data, and
- singling out protection across **multiple disclosures**.

In particular, in statistics, a minimal cell size is prescribed. Only data that are aggregated in each cell over a minimal number of persons[121] is then permissible. In contrast, the concept of anonymization technique also includes methods such as K-Anonymity that attempt to create anonymous individual-level information. Disclosure control[122] is stricter by disallowing singling out in anonymized data and thus excludes the possibility of unaggregated individual-level anonymous data.

Also, statistical disclosure control goes beyond just "anonymizing" a single original data set. Much rather than preventing that individual-level information can be extracted form a single anonymized data set, it is concerned with doing this for a **multitude of "anonymized" data**. These are typically referred to **disclosures**. While one disclosure may be insufficient to gain information about individuals, multiple disclosures may well be combined to gain such information. Using multiple disclosures to partly invert the aggregation is typically called **reconstruction attack**. In contrast to disclosure control, "anonymization techniques" may be concerned with only a single disclosure and thus disregard the possibility of reconstruction.

It is important to understand that the singling out protection of disclosure control goes far beyond just aggregation. **It is a common misconception that aggregation protects against singling out**. This is only true for a single disclosure. Reconstruction attacks that use multiple disclosures have been prominently demonstrated by the U.S. Census Bureau[123]. They have also been shown in DataShield although it can limit single disclosures to a statistics over a minimal cell size[124] [125].

Aggregation implements the idea of "blending into a crowd" (a term originally proposed by Reiter and Rubin[126]. This is much easier for an average person than for one who is exceptional or

[120] See for example Anco Hundepool et al., Center of Excellence SDC, Handbook on Statistical Disclosure Control, Second Edition, November, 2024, https://sdctools.github.io/HandbookSDC/Handbook-on-Statistical-Disclosure-Control.pdf (last visited 16/12/24).

[121] In statistics, the aggregation may also be over households or enterprises.

[122] The use of the term "disclosure control" in this document implies going beyond just basic aggregation. It is consistent with the terminology presented in D4.9.4. Note that in contrast, some authors use the same terms as synonym for "identity-reduction", i.e., any transformation that reduces the risk of re-identification. For example, the term disclosure control was even applied to microdata [https://link.springer.com/book/10.1007/978-3-319-50272-4] in total absence of aggregation.

[123] 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/10.1145/3287287 (last visited 16/12/24).

[124] https://datashield.discourse.group/t/vulnerability-in-federated-analysis-software/622 (last visited 16/12/24).

[125] https://datashield.discourse.group/t/statement-datashield-disclosure-controls-and-mitigation/628 (last visited 16/12/24).

[126] Michael K. Reiter and Aviel D. Rubin. 1998. Crowds: anonymity for Web transactions. ACM Trans. Inf. Syst. Secur. 1, 1 (Nov. 1998), 66–92. https://doi.org/10.1145/290163.290168

outstanding in certain ways. An illustrative example is that it would be difficult to hide a single millionaire living in a census district that is otherwise dominated by slums. Further evidence is that high-dimensional data is commonly acknowledged as very difficult to anonymize. The reason for this is that a person may blend in well when considering few properties, but is usually outstanding when considering many. In other words, when considering a small number of properties, a person is likely to have several "close neighbors"; when considering a large number of properties, the distribution gets much sparser and a person may well lack any "close neighbor"[127].

For this reason, disclosure control typically encompasses to identify **data subject with outstanding properties** in order to either **suppress** their data (since just unprotectable) **or** make them focus of an **increased level of protection** (such as increased randomization or noise). Similarly, in differentially private anonymization techniques, the impact of exceptional values can be reduced[128] since their full inclusion would require a much higher level of noise than is otherwise necessary.

An illustrative example of detecting outstanding individuals is the **Special Uniques Detection Algorithm** (*SUDA*)[129] [130]. It is implemented in an easy to understand open source python package[131]. While the original algorithm is limited to recognize neighbors based on the equality of data values (i.e., defined for categories or integers), the method has also been generalized to working based on closeness of continuous values[132] [133].

5.4.2 Preventive Anonymization

How exactly to perform an effective anonymization is not a clear cut task. Much rather, anonymization techniques typically have a large number of parameters that need to be set. These include for example decisions on how to reduce detail (e.g., how to select intervals of values), which values or data subjects to suppress (i.e., a selection of thresholds), and how much noise or randomness to inject where.

Often, the determination of anonymization parameters requires a recursive approach. For example, in K-Anonymity, value intervals are chosen in a first step, then evaluated in terms of the resulting minimal size of equivalence classes, and then when refined if necessary. Similarly, in statistics, one may choose a certain geographic aggregation strategy and then evaluate and refine it depending on the size of the resulting cells.

Considering that careful anonymization is thus likely a time consuming task, anonymization cannot be performed on the fly for every single request. This is further supported by the reason that

[127] This could also be seen as clustering. Average persons who blend into a cluster of close neighbors are easier to hide in the crowd than outstanding persons.

[128] See for example Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, Li Zhang, *Deep Learning with Differential Privacy,* ACM CCS 2016, https://arxiv.org/abs/1607.00133 (last visited 13/12/24).

[129] 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.

[130] Elliot, M. J., Manning, A., Mayes, K., Gurd, J., & Bane, M. (2005). SUDA: A Program for Detecting Special Uniques. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Geneva.

[131] https://pypi.org/project/suda/ (last visited 16/12/24).

[132] Ichim, Daniela. (2009). Disclosure Control of Business Microdata: A Density-Based Approach. International Statistical Review. 77. 196-211. 10.1111/j.1751-5823.2009.00079.x.

[133] D. Ichim, Microdata anonymisation of the Community Innovation Survey data: a density based clustering approach for risk assessment, DOCUMENTI ISTAT, n. 2/2007, https://www.istat.it/wp-content/uploads/2018/07/2007_2-1.pdf (last visited 16/12/24).

anonymization can often be more effective if large input data sets are available. In contrast, requests for anonymous data may be geographically limited and thus reference only small data sets.

A common way to implement this approach is to generate synthetic data use these to create responses to requests. This approach has the additional advantage that the very same tooling can be used as those for answering requests from the original data. In the scenario shown in Figure 4, complete snapshots could be the unit for an anonymization[134]. Further, since the time consuming anonymization (that may ever require human involvement) is performed before any requests, it becomes easier to automate the processing of requests and perform them on the fly. This would result in request processing without human intervention but still guarantee the quality of the anonymization that may require assessment by humans.

5.4.3 Double protection against re-identification

Another measure to reduce the risk of re-identification is to use two kinds of protection. Typically, a first protection results in still individual-level data, while a second adds further protection through aggregation that may optionally be further protected against reconstruction (e.g. through noise injection). This approach provides a second line of defense: Even in the case that reconstruction should be possible, the reconstructed data is still protected to a certain degree.

An example for this approach is to first use generalization (e.g., exact times to days, precise coordinates to postal zones) and suppression (or top coding) of outstanding data values (e.g., delete records with very high values) to yield k-anonymous individual-level data. These are then aggregated, for example with statistics in a second step. Here, even if reconstruction of the individual-level data is possible, k-anonymity still protects (yet to a lesser degree) against re-identification. Further, particularly easy to re-identity outstanding data subjects have been eliminated and are no longer at risk

Another example is to create synthetic data in a first step and aggregate them in a second step. The first step of generating synthetic data is itself built on an aggregation. Namely, a form of aggregation is used to yield a (statistical or machine learning) model from which synthetic individual-level data are then randomly generated. In this approach, even if the aggregation can be inverted through reconstruction, attackers can only obtain synthetic data that still protect data subjects against re-identification.

5.4.4 Using anonymization with strong guarantees

Some anonymization techniques like epsilon differential privacy provide strong guarantees against re-identification. These are partially based at assuming the worst case or attackers' background knowledge, allowing arbitrary post-processing, and being composable across multiple disclosures.

In contrast, methods that cannot offer such guarantees always rely on certain assumptions in order to guarantee anonymity. Examples for such assumption is that linkable additional information is unavailable or that no known attack against the protection mechanisms exist, or that computational

[134] One counter argument would be if the snapshot had a too restricted data volume. This could for example be the case when considering rare diseases in the EHDS. In this case, a geographic (horizontal) integration of the snapshots of multiple data holders before anonymization may be a reasonable strategy.

cost is excessive. Evidently, such assumptions can change in time. In particular, new additional can become available[135], new attacks can be found, and computing power can increase.

For this reason, a periodic review of anonymized data is often recommended[136]. In particular, the responsible parties need to establish whether the assumption still hold[137]. Anonymization methods that offer strong guarantees largely ease this burden. In particular, the only factor that can threaten the continuing efficacy of the anonymization are additional disclosures. In differential privacy, the loss of efficacy is called erosion of the privacy budget.

In a data space, the disclosure of anonymized data continuously increases over time. This is illustrated in Figure 19. If it is necessary to periodically verify the efficacy of disclosed anonymized data, the burden for the responsible party continuously increases with the data. The resources available to this party thus could limit the accumulative volume of anonymized data that is economically manageable in a data space. It is evident that anonymization methods with strong guaranteed drastically ease the burden on responsible parties and thus drastically increase the manageable data volume.

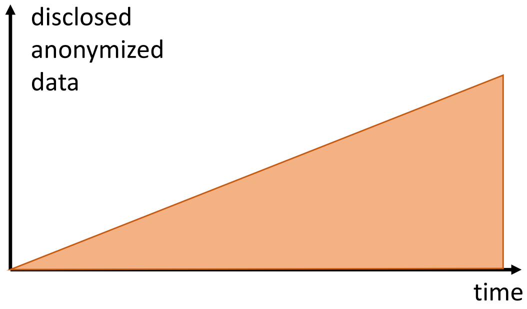


Figure 19: Disclosure of anonymized data over time.

5.4.5 Strategy to delay the erosion of strong guarantees.

This subsection first provides the necessary background on how strong guarantees erode with every new disclosure of anonymized data and, on this basis, describes a strategy to delay the erosion of guarantees.

A common understanding in the scientific community is that every disclosure of anonymized data leaks at least some information and that there is the danger of accumulating such information and use it to (at least partially) break the anonymization.

[135] Note the development of ever more ubiquitous networked sensors as for example provided in smartphones or fitness watches. Another example that makes additional information available is a data breach concerning primary use data.

[136] Note that this is incompatible with a "publish and forget" approach.

[137] The Article 29 Data Protection Working Party states in 0829/14/EN, WP 216, Opinion 05/2014 on Anonymisation Techniques, Adopted on 10 April 2014, that "Thus, anonymisation should not be regarded as a one-off exercise and the attending risks should be reassessed regularly by data controllers."

In statistics, such information leakage can be used to construct big equation systems of reconstruction attacks[138] and (at least partially) reverse the aggregation to single out data subjects and obtain individual-level information.

The U.S. Census Bureau has evaluated the risk of such reconstruction for their 2010 Census data. The 2010 census used disclosure control and thus incorporated measures to protect against reconstruction (namely[139] cell suppression for small groups, top-coding for age data, noise-injection for some attributes, and swapping of attributes). In spite of these measures, they successfully performed a reconstruction attack and were able to re-identify a significant percentage of data subjects by linking to additional information they could purchase commercially[140].

Considering the unexpected success of the reconstruction attack, they decided to base the 2020 Census on differential privacy that offers strong guarantees against reconstruction. Differential privacy is based on systematic and calibrated noise injection. Every disclosure can be seen as incorporating an independent realization of random noise.

In differential privacy, accumulation a sufficient number of disclosures therefore enables attackers to "average out" the noise to yield a picture with much reduced or no noise. Evidently, less noise implies less protection against re-identification. The protection provided by differential privacy is usually expressed as ***privacy budget*** (aka. epsilon). This privacy budget erodes with every new disclosure[141]. It is important to realize that eroding disclosures could originate from multiple controllers who may not coordinate their disclosures or even know of one another.

For data spaces to be viable over time, the erosion of strong guarantees needs to be managed by the disclosing parties. Since the disclosed anonymized data is considered anonymous, it is likely that such data is also made public (directly by the disclosing party or indirectly by the recipient). Where a high volume of disclosure is expected, the privacy budget may erode rapidly and render reconstruction attacks possible. A large-scale successful attack in a data space may well destroy trust by participating citizens and seriously endanger the data space itself.

The management of the privacy budget (i.e., the erosion of strong guarantees) is therefore an important requirement for data spaces. How to best achieve this in data spaces where the law dictates an unlimited number of disclosures is likely challenging. This can be further aggravated by the continuous evolution of the underlying primary use data. The following falls short of suggesting a comprehensive solution but shows a strategy to delay the erosion of the privacy budget.

The strategy consist of anonymizing the full data set once and answer many request from the anonymized data set. For example, the anonymization could create differentially private synthetic data[142]. The results of requests is then also anonymous. This is shown in Figure 20. It shows that

[138] Garfinkel, Simson & Abowd, John & Martindale, Christian. (2019). Understanding database reconstruction attacks on public data--These attacks on statistical databases are no longer a theoretical danger. Communications of the ACM. 62. http://dx.doi.org/10.1145/3287287 or https://queue.acm.org/detail.cfm?id=3295691 (last visited 9/7/24).

[139] See footnote above.

[140] In Europe, likely such a market for suitable additional information does not exist. But the findings of the U.S. Census Bureau are equally applicable.

[141] This is due to the independent noise realization of each disclosure and the fact that the noise follows a distribution centered on the true values.

[142] Synthetic data are artificial data that are generated from a model. This model describes the original data. For example, a model may capture the statistical distribution of original data and then use this the distribution model to generate random data points. Similarly, a machine learning model could be trained on the original

only a single noise realization is used across all requests. This avoids that a high volume of requests rapidly erodes the privacy budget.

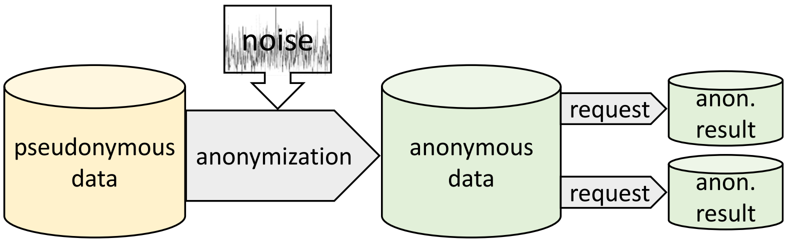


Figure 20: Strategy to anonymize once in order to answer many requests.

This strategy contrasts with one where every result of a request is anonymized individually.   This is shown in Figure 21. It shows how every anonymization requires its own noise realization. In case of a high volume of requests, this leads to a rapid erosion of the privacy budget and thus strongly guaranteed protection against reconstruction. In particular, in this case, a reconstruction attack can be structured in a way that the effect of the added noise can be averaged out across many noise realizations.

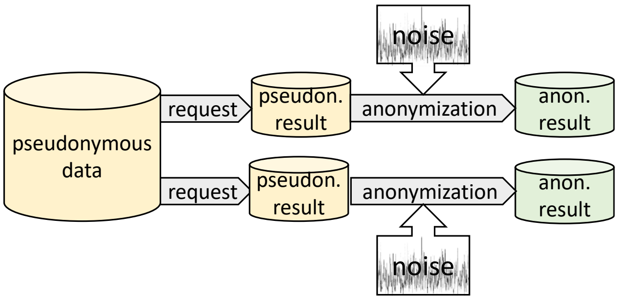


Figure 21: Strategy to anonymize the result of every request individually.

In data spaces where continually, new data (partially about already existing data subjects) are added to the primary use data, using the same noise realization in many snapshots seems to be impossible. This scenario is visualized in Figure 22. Here, the anonymization of every snapshot erodes the privacy budget due to an independent noise realization. Since the frequency of snapshots can be expected to be orders of magnitude lower than that of requests, the shown strategy at least delays the erosion.

data.   It is then used to generate random data points with very similar characteristics (such as correlations between and distribution of attributes).

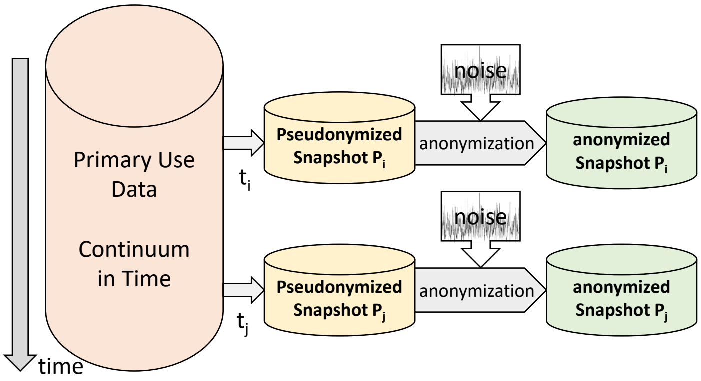


Figure 22: Strategy to anonymize every (entire) snapshot drawn from a continually changing primary use data set.

Synthetic data represent an attractive approach for creating anonymized snapshots of data. The following technically detailed description proposes an evolved strategy in more detail.

5.4.5.1 Proposed strategy to create anonymous snapshots in a data space[143]

Preserving privacy in the secondary use of the EHDS-data is an essential regulatory requirement. Another central requirement to the secondary use of EHDS-data is the reproducibility of studies. Moreover, there is a subtle fairness challenge when dealing with privacy-preserving usage of sensitive data: Due to likability attacks, every additional study on the same sensitive data leaks more information about that data. One direction for solving this problem is to be more strict with each usage of the data in order to ensure that future usages of the same data does not lead to too high privacy leakage, and thus a privacy violation. Concretely, being more strict in terms of privacy protection could mean requiring larger micro-aggregation clusters or using more noise when processing the data.

The research literature suggests that more accurate results can be achieved when processing techniques are redesigned to preserve privacy (e.g., Differentially Private Data Processing). However, such adapted processing techniques does not necessarily meet the fairness requirements: subsequent users can benefit from information extracted by previous studies, while earlier users may not always be able to improve their studies using the results of later studies. Additionally, privacy-preserving algorithms present challenges for reproducibility that do not arise in the case where the used data can be directly inspected by future researchers.

A future-proof design for the secondary use of EHDS data could employ regularly updated data synthesis to facilitate fairness and reproducibility. Privacy-preserving data synthesis, however, comes with its own challenges. Data synthesis is a harder task and might not lead to the same accuracy as if the data processing algorithms would be redesigned. In particular, a synthetic data cannot retain all information from the original dataset. Consequently, various approaches exist to balance the utility of synthetic data and the protection of the patients' privacy. As a remedy, we propose the following approach:

1. Provision of Synthetic Data with Quality Guarantees: EHDS operators provide a set of synthetic datasets for each original dataset. Users can freely use and refer to any of these synthetic datasets for their studies. This ensures reproducibility. Additionally, EHDS operators supply utility guarantees for each dataset, characterizing which types of information (e.g., correlations between attributes) are preserved and to what extent.

[143] This subsection was contributed by Esfandiar Mohamadi of UzL-Privsec.

2. On-Demand Dataset Synthesis: If the quality guarantees of existing synthetic datasets are insufficient for the users (e.g., correlations between attributes A and B are not preserved), users can register their studies and request the synthesis of a new dataset that provides better utility guarantees for their study (e.g., correlations between attributes A and B are preserved within a 5% margin of error).
3. Periodic Release of Synthetic Data: At regular intervals (e.g., once per month), the EHDS operator publishes new synthetic datasets along with corresponding quality guarantees. In this process, the operator considers all submitted studies and synthesizes the data to meet the quality requirements of these studies.
4. Re-Running Previous Studies: The EHDS operator reruns all prior studies on the newly released synthetic datasets and informs the respective study operators of the updated results. This procedure ensures fairness for earlier users.

Through this approach, sensitive patient information can be efficiently utilized while maintaining controlled disclosure of sensitive data derived from patient records.

5.4.6 Verification Servers to extend the usability of anonymous data

Anonymization usually introduces an artificial bias in the data. This bias results from the deviation to truthfulness (e.g., noise injection), suppression of data relating to outstanding subjects, and reduction of detail[144]. While from a data protection point of view, it is much preferable to use anonymous data, one reason to use pseudonymous data instead is that it remains unclear whether results of an analysis are indeed significant or are caused by an artificial bias that was introduced by the anonymization. The present subsection describes the concept of verification servers that solve this problem.

In particular, verification servers provide a quality measure that expresses how a result based on anonymous data compares to the same analysis run on truthful, complete, and detailed pseudonymous data. In case that an analysis of anonymous data yields results of sufficient quality, it becomes unnecessary to access the equivalent pseudonymous data.

A verification server can thus eliminate the need to access pseudonymous data in many cases. This fails to hold for cases where an insufficient quality is found. This indicates that the result from only anonymous data is degraded by a significant bias. This is a strong and easily documentable reason to request access the equivalent pseudonymous data in an SPE. It may even be an efficient policy to request to conduct an analysis on anonymous data first as a prerequisite for requesting access to pseudonymous data. Such a strategy is enabled by verification servers.

The concept of verification server was proposed by Reiter et al[145] [146]. The concept is visualized in Figure 23. It shows how a truthful, complete, and detailed pseudonymous version of the data co-

[144] As an example for how generalization can introduce a bias, consider that precise locations are mapped to towns. If points are located on both side just along the boundary of two towns, the generalized data may suggest a significant distance between the points. This would be incorrect and a bias introduced by the design of generalization.

[145] Reiter, Jerome & Oganian, Anna & Karr, Alan. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. Computational Statistics & Data Analysis. 53. 1475-1482. 10.1016/j.csda.2008.10.006.

[146] Karr, Oganian, and Reiter, Verification Servers, in Int. Statistical Inst.: Proc. 58th World Statistical Congress, 2011, Dublin (Session IPS060), https://2011.isiproceedings.org/papers/450140.pdf (last visited 18/12/24).

exists with a likely biased anonymized version (for example, in the form of synthetic data). Data users then request to perform a certain *analysis procedure* on the data. This is typically expressed in form of an executable script or program. The *verification server* then executed this analysis procedure on both, the pseudonymous and the anonymous data. This yields two results, one on the pseudonymous side that cannot be guaranteed to be anonymous, and a second one on the anonymous side that is indeed guaranteed to be anonymous[147]. The verification server now compares the two results and summarizes the difference in a generalized *quality assessment*. To preserve anonymity, this quality assessment must be much less detailed than providing actual differences in values. For example, it could provide statistics about differences or, depending on the kind of analysis, just provide a single assessment of significance on an ordinal scale. Data users then receive the anonymous result of the analysis together with its quality assessment.

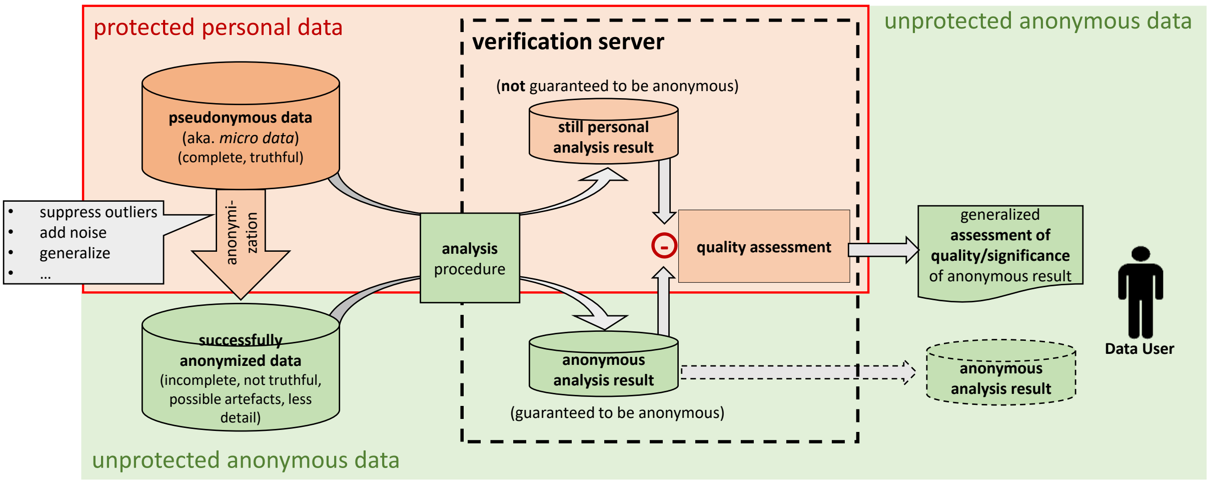

protected personal data
verification server
unprotected anonymous data
pseudonymous data
(aka. *micro data*)
(complete, truthful)
(not guaranteed to be anonymous)
still personal analysis result
suppress outliers
add noise
generalize
...
anonymi-zation
analysis procedure
-
quality assessment
generalized assessment of quality/significance of anonymous result
Data User
successfully anonymized data
(incomplete, not truthful, possible artefacts, less detail)
anonymous analysis result
(guaranteed to be anonymous)
anonymous analysis result
unprotected anonymous data

Figure 23: The concept of verification servers.

Evidently, in this scenario, all analysis procedures that yield a good quality assessment can be limited to exclusively disclosing anonymous data to data users. In contrast, analysis procedures with bad quality assessment justify that data users apply for access to pseudonymous data.

It can be followed from this consideration that verification servers can significantly reduce the data protection risk of a data space by reducing the need for disclosure of pseudonymous (i.e., personal) data while maintaining the same functionality (i.e., reply to the same analysis requests).

In more detail, it may also enable data users to evolve their analysis purely on anonymous data. The final version of the analysis is then captured in an executable form and run through the verification server. This can further reduce the exposure of personal data in a data space.

5.5 Step 5: Publishing anonymized results of analysis conducted in an SPE

Data users who conduct an analysis of pseudonymous data in a secure processing environment (SPE) typically publish an anonymous version of their results. This is shown as step 5 in Figure 4. The motivation for publishing results could for example be legal or scientific. An example of the former case is a legal obligation for data users as expressed in Art. 46(11) EHDS proposal. Namely, it reads: “Data users shall make public the results or output of the secondary use of electronic health data,

[147] This guarantee holds at least if the underlying data set was successfully anonymized.

[…]" and "Those results or output shall only contain anonymised data". The common practice of publishing in the science community provides a motivation for publishing results even where a legal obligation is lacking.

Based on the background provided in the discussion of the previous step, this section discusses how to publish anonymized results obtained from the analysis of pseudonymous data in an SPE.

The most intuitive approach of directly anonymize results obtained in an SPE. For future data spaces, it is not yet clear under whose responsibility the anonymization falls[148]. It is assumed here that the responsibilities are shared between the provider of the SPE and data users as follows.

Providers of the SPE are responsible for:

- Issuing policy and guidance on how to anonymize results,
- assessing that the level of anonymity is sufficient before approving that results and outputs leave the SPE.

Data users are then responsible for:

- Performing the actual anonymization according to policy and guidelines.

This intuitive approach is very problematic, however. Among the issues are the following:

- The fact that SE providers have to approve the anonymized result of every analysis can proof to be onerous (depending of the volume of requests and available resources). This could be further aggravated by legally prescribed maximal response times.
- Leaving the task of the anonymization to data users can create heterogenous standards of anonymization and levels of protection. This can already be the case within a single SPE but is more likely to occur at European level.
- Data users are specialist in their field of endeavor but may lack knowledge and experience in technical and legal aspects of anonymization[149]. This may constitute an undesired hurdle for data users or an increased burden for SPE providers.
- Finally but most importantly, the focus of anonymization is limited to a single disclosure. In contrast, the scope of disclosure control encompasses multiple disclosures, in this context the analysis results of all data users. With the understanding that every disclosure inherently leaks some information, an unlimited number of data users and requests seems to indicate a very rapid erosion of the privacy budget. Leaving data users responsible for anonymization further aggravates the difficulty of SDE providers to implement full disclosure control.

[148] The EHDS proposal states in its Article 51(1) that "The health data access bodies and the data users, […], shall be deemed joint controllers […]".

[149] The *Opinion 05/2014 on Anonymisation Techiques* (WP 216) by the *Article 29 Data Protection Working Party* seems to indicate a level of complexity that likely leads to a significant learning curve for the topic of anonymization. This assessment is further supported by the current lack of commonly accepted standards, procedures, and policies concerning sufficient anonymization, as well as the difference in understanding in the scientific and legal community. (For WP216, see https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf)

A possible solution for the described issues takes a very similar approach to that of verification servers describe in Section 5.4.6 together with preventive anonymization of Section 5.4.2. This is illustrated in Figure 24.

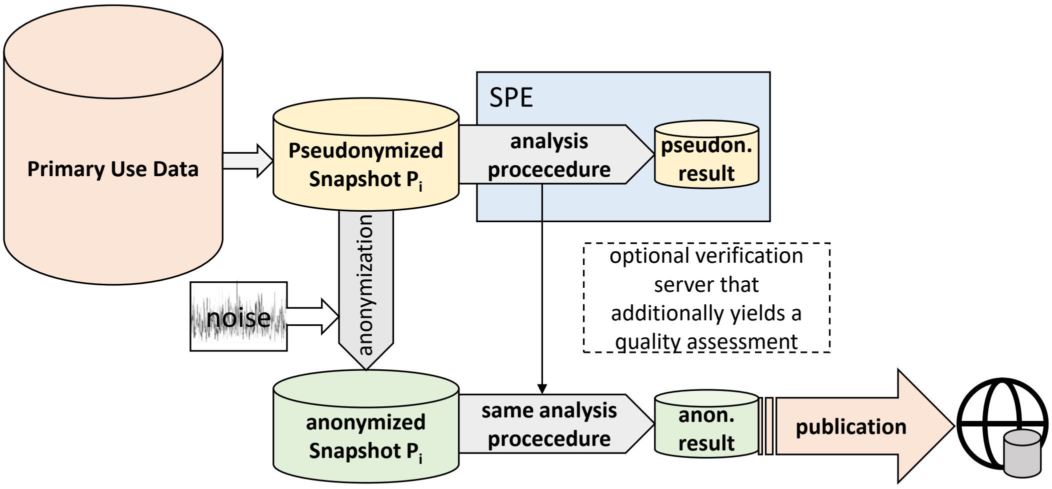

Primary Use Data
Pseudonymized Snapshot P_i
SPE
analysis procecedure
pseudon. result
anonymization
noise
optional verification server that additionally yields a quality assessment
anonymized Snapshot P_i
same analysis procecedure
anon. result
publication

Figure 24: Proposed approach to publishing analysis results from an SPE.

The approach assumes that an effective anonymization of an analysis result can be difficult to achieve. In particular, even if the results consist of only statistics, typical elements of effective anonymization (namely reduction of detail, suppression of outstanding data subjects and values, injection of noise) are not incorporated in the anonymization. The re-identification risk of such disclosures must thus be considered to be relatively high, particularly together with a multitude of related disclosures. For this reason, results of an analysis of truthful, complete, and detailed pseudonymous data were themselves considered to be pseudonymous. Pseudonymous results can obviously not be allowed to leave the SPE though publication.

Assuming the implementation of preventive anonymization, an effectively anonymized data set exists that corresponds to the pseudonymous one which is accessed through the SPE. To yield an anonymized result with more confidence, the same analysis procedure[150] that generated the result in the SPE can be applied to the preventively anonymized data. This yields anonymous data that can then be published.

As was discussed in Section 5.4.5, the proposed approach permits the SPE provider to better control the overall privacy budget across all analysis results from many data users. This is possible though using only a single noise realization during preventive anonymization, independently of the number of analysis results.

[150] Note that this does not necessarily imply that the analysis has to result in a single executable script of program. Since SPEs log all activities by data users for other reasons, a simple replay of the log could be used. This lacks efficiency since there may be "dead branches". The efficiency problem could probably be fixed by manually or automatically pruning dead branches from the activity log.

6 Summary and Conclusions

Europa is currently implementing a strategy to foster large-scale secondary use of data in sectorial data spaces. Since the strategy also concerns personal data, the legal data protection requirements of the GDPR have to be satisfied. Since data spaces can be complex and multi-faceted, how to best incorporate data protection needs careful consideration.

This document aims to contribute with a first analysis of this problem. In particular, it has discussed the most relevant data protection requirements in the context of data space and pointed out where data protectors see the major risks. It has then developed a simple model for data spaces first looking at main characteristics and then describing its functionality in five major steps. Based on a survey of the literature, it has then described possible technical and organizational measures in support of data protection for each step.