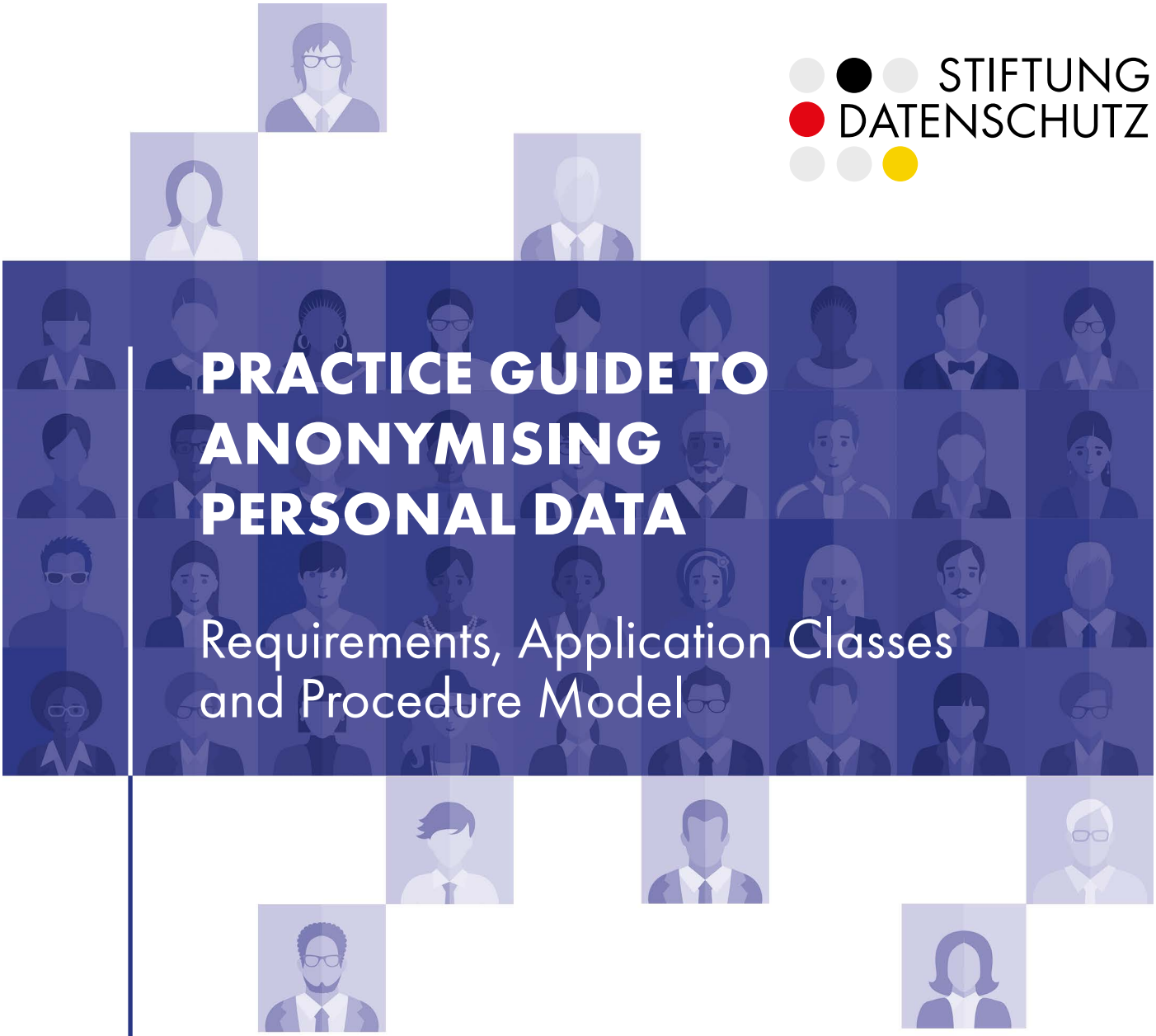
STIFTUNG DATENSCHUTZ



PRACTICE GUIDE TO ANONYMISING PERSONAL DATA

Requirements, Application Classes and Procedure Model

AUTHORS

Prof. Dr. Rolf Schwartmann, Andreas Jaspers, Dr. Niels Lepperhoff, Steffen Weiß LL.M., Prof. Dr. Michael Meier

Practice Guide to Anonymising Personal Data

Requirements, application classes and procedure model

prepared and submitted on behalf of the Foundation for Data Protection

in December 2022

from

Professor Dr Rolf Schwartmann
Cologne Research Centre for Media Law, Cologne University of Applied Sciences,
German Association for Data Protection and Data Security (GDD) e.V.,

Andreas Jaspers, attorney-at-law
German Association for Data Protection and Data Security (GDD) e.V.,
DSZ Datenschutz Zertifizierungsgesellschaft mbH,

Dr Niels Lepperhoff
DSZ Datenschutz Zertifizierungsgesellschaft mbH,
XAMIT Bewertungsgesellschaft mbH,

Steffen Weiß, LL.M., attorney-at-law
German Association for Data Protection and Data Security (GDD) e.V.

with the assistance of

Professor Dr Michael Meier*
University of Bonn, Fraunhofer FKIE,
Society for Data Protection and Data Security (GDD) e.V.

*With the participation of Saffija Kasem-Madani, Markus Krämer, Daniel Meyer (all University of Bonn)

Table of Contents

1. Introduction

Data is at the heart of the digital transformation. The exchange of personal data between public and private entities is steadily increasing. The European Data Strategy aims to help the EU become a leader in the international competition for data-driven business models.

The aim of the GDPR is to put the processing of personal data at the service of humanity. This includes purposes of the common good as well as business interests. The GDPR aims to **enable, not prevent,** data processing in accordance with practical and economic needs. In this respect, data protection is, to use an image of connected driving, more of a lane assistant than a brake. The GDPR is open to new data-driven economic and technical developments. It does not stand in the way of new contractual constructions and the associated data processing, even under the heading of "numbers with data". Services designed to maximise data, such as social networks, cannot be legally understood in any other way. After appropriate consideration, the further processing of data for new purposes that are compatible and in line with interests is also permitted.

Data protection law by no means only allows for the processing of anonymous data that do not relate to individuals at all. In particular, **encryption and pseudonymisation of** personal data serve as **"enabling tools"** for the technical safeguarding of protection when processing even large amounts of data. The Federal Government's Data Ethics Commission has given important space to the topic of data access for both personal data and non-personal data in sections E 4. and E 5. [1]

Politically, the data protection-compliant evaluation of health data of diseases is particularly desired. In spring 2022, the European Commission presented a draft regulation to create a European health data space. It is intended to give individuals control over their health data and at the same time open up the use of health data for better medical care and research. The EU should exploit the potential of health data exchange, use and re-use. The Federal Government's Data Ethics Commission also advocates this under 3.5.2 of the final report.

The coalition agreement of the German coalition government has taken this up. The further development of the **digitalisation of the health system** has been agreed. Rightly so, because the fight against the pandemic at the latest has shown us how important it is to create a balanced framework for the processing of health data that puts the concerns of health protection in an appropriate relationship to the protection of privacy.

In addition, there are other practical use cases of anonymisation, from the use of digital street maps, to aggregated user statistics in the online sector or in the context of contractual customer relationships.

[1] Final Report of the Data Ethics Commission (2019).

This is where the basic rules of this study commissioned by the Foundation for Data Protection on **anonymisation** come in. The subject of this work on anonymisation is intended to be the **basis for area-wide applications** in practice. Anonymisation of data makes it impossible to draw conclusions about a person and excludes them from the scope of the GDPR. This is desired and necessary when anonymous data fulfils the purpose after processing. If data subjects want to be able to draw conclusions about their data, for example for their own health care, the GDPR provides for pseudonymisation. It ensures that data is protected against misuse through encryption. In this regard, the study refers to the **"Draft for a Code of Conduct on the use of GDPR compliant pseudonymisation"** prepared by the Focus Group on Data Protection of the Federal Ministry of the Interior within the framework of the Digital Summit 2019.

The linchpin of the legal scope of application is the question of the reference to persons. If this is removed, the legal requirements cannot claim any effectiveness. According to the considerations of the GDPR, data protection should apply to all information that relates to an identified or identifiable natural person. If these characteristics are not present, a datum is anonymous. For the further use of personal data, it is significant that the law also considers the anonymisation of personal data as possible (see recital 26). After the transformation or alteration of personal data, the data subject cannot be identified or can no longer be identified.

The desire of data controllers to leave the scope of the GDPR by means of anonymisation in order to make data more easily usable by themselves or by third parties is understandable. However, it is all the more important to clearly draw the line of what is legally permissible, first of all with regard to its basic rules. This is the only way to create legal clarity and legal certainty while complying with the requirements of the GDPR.

This Practice Guide deals with information on the anonymisation of personal data.[2] For this purpose, the **term anonymisation** and its characteristics are classified in the existing legal context. A distinction must be made from other processing operations, namely pseudonymisation. After the conceptual classification, common **procedures and techniques of anonymisation** are described in general. In order to maintain the practical relevance, this is followed by application classes of anonymisation. Here, users are shown **application scenarios and examples** in which anonymisation can take place. A separate chapter will deal with the legal environment of anonymisation and the existing requirements, be it special testing, documentation or transparency obligations. In order to **support small and medium-sized enterprises** in particular, a procedure model is provided to carry out the process of anonymisation step by step and in a structured manner.

The following description serves as a general guide for the anonymisation of personal data. It cannot and should not serve to formulate conclusive specifications for such a procedure. In addition, other laws may contain requirements for non-personal data that apply accordingly to an anonymised data set, such as the Data Governance Act.

[2] On the requirements for pseudonymisation Schwartmann/Weiß, Anforderungen an den datenschutzkonformen Einsatz von Pseudonymisierungslösungen - ein Arbeitspapier der Fokusgruppe Datenschutz der Plattform Sicherheit, Schutz und Vertrauen für Gesellschaft und Wirtschaft im Rahmen des Digital-Gipfels 2018.

2. The Practice Guide at a glance

2.1 Framework conditions for anonymisation

Anonymisation of personal data is a complex process. First of all, it requires a basic understanding of central **concepts** with a view to their framework conditions**.** This starts with the basic understanding of data, which can exist in structured as well as unstructured form**.**

- **Structured data** can be understood as key-value pairs, while **unstructured data** is any data that does not conform to a key-value pair representation.
- **Personal data**, as a further manifestation of data, are defined by the GDPR and have a wide scope of application due to the legislative reference to "identifiability".
- While **anonymisation** aims to remove a reference to a person, re-identification of data subjects is still possible within the framework of **pseudonymisation.** However, the data that can be used for re-identification must be stored separately and protected by technical and organisational measures.
- The **transition** between anonymisation and pseudonymisation can sometimes be **fluid.** Especially when processing takes place in a different context and additional information is available that contributes to the identification of a data subject.

2.2 Terms of anonymisation with the inclusion of valuations

There are various approaches to **anonymisation**.

- **Absolute anonymisation** is characterised by the fact that the re-identification of then person concerned is completely excluded.
- **Factual or relative anonymisation is** characterised by the **fact** that the re-identification of the person concerned is not completely excluded. However, re-identification of the data subject is ruled out due to the **disproportionate effort involved.**

In its recitals, the GDPR opens up the possibility of including proportionality aspects in the question of successful anonymisation. The **case law of the highest courts** does not consider it necessary that all information necessary for identification is in the hands of a single controller. Rather, it is sufficient if the controller has the data subject identified via a third party. The limit of identifiability again lies in the impossibility, disproportionality or violation of the law.

2.3 Special case: Artificial intelligence and related techniques

The discussion of anonymisation in the context of **artificial intelligence (AI)** requires an understanding of the background of this newer technology. Basically, AI is nothing more than such a computer programme. There are many different types of **algorithms and procedures**, each of which functions differently and has different areas of application. One of the most dominant classes of algorithms in the public eye today is pattern recognition using machine learning. Pattern recognition algorithms contain non-specific rules that are only adapted in a "training phase". Algorithms and a result determined from them are also subject to the demarcation difficulties between personal and anonymous data.

2.4 European legal context: GDPR and "new data file

2.4.1 New data file

Anonymisation of personal data is not only in a technical but also in a legal environment. The Commission's **European data strategy** targets newer technologies, for example, via the draft AI Regulation, as well as data sharing via the Data Governance Act and the draft Data Act. This is intended to create new **data spaces** for certain sectors and create legal incentives such as obligations for the sharing of both personal and non-personal data. The mechanisms for data sharing provided for in the new legal acts mostly offer the possibility of **implementing protective measures for the** benefit of data subjects. These include **pseudonymisation** and **anonymisation** of personal data. The anonymisation of data will therefore gain significantly in importance with the new data acts. The GDPR remains the standard when it comes to processing personal data, which must be complied with in addition to specific requirements of the new legal acts.

2.4.2 GDPR

As an existing standard for anonymisation, the **GDPR** requires a **legal basis** for the anonymisation of personal data. Since the purposes of an initial collection of personal data and those of anonymisation regularly differ, it must be examined whether these **purposes are compatible**. Such compatibility can usually be assumed if the anonymisation is not based on special categories of personal data.

In principle, anonymisation leaves the scope of the GDPR. In some cases, however, there are uncertainties as to whether successful anonymisation can be assumed according to the legal standards. Controllers are not prevented from understanding anonymisation as a **technical-organisational measure** that is applied to personal data. This means that a data controller is still within the legal scope of application, but he or her can take credit for the protective measures taken, for example, in the context of a balancing of interests.

The **requirements** for anonymising personal data can basically be categorised into a **legal and a technical part.** However, since a third party's interest in anonymised data is of particular importance for the choice of technical means of anonymisation, the description of an **attacker model** is listed as a further requirement by the Practice Guide.

From a legal point of view, there is a **duty to verify** whether the data generated from anonymisation relates to an identified or identifiable individual. Within the framework of this examination, **all means** must be taken into account that could reasonably be used either by the controller or a third party to identify the data subject. Such means may be, for example, information available to the controller or information that the controller can obtain. Possible linkages of data must also be included in the examination. Especially in view of the **European Data Strategy,** publicly accessible data rooms with personal, pseudonymised or anonymised data are increasingly to be expected. A variety of data sources can increase the probability of re-identification. It is irrelevant whether the controller or a recipient wants to identify data of the individual or not. Objective identifiability is sufficient.

From the perspective of the GDPR, in the absence of legal specification, various anonymisation techniques can be used. The decisive factor is that, after examining the factors listed above, re-identification of data subjects is not practically feasible. I.e. if it requires a **disproportionate effort in terms of time, costs and manpower,** effective anonymisation can generally be assumed.

2.5 "Attacker model" as a test

An **"attacker model"** describes a method to check whether a data set is anonymous or personal. From the perspective of an attacker, it is tested whether re-identification is possible. The knowledge and skills assumed of the attacker depend on the context of use of the data, for example, whether anonymised data is published, only used internally or passed on to specific recipients. It is advisable to consider not only targeted attacks in the attacker model, but also constellations in which re-identification by the attacker is actually unintentional or could occur by chance. There are various ways to carry out an attack with the aim of re-identification. These include the **singling out ~~of~~ a person,** the **linking of data records (record linkage)** and the derivation of characteristics of a person from other characteristics available in the database **(inference)**.

2.6 Anonymisation methods

From a **technical point of view**, various anonymisation methods are available, which can be divided into those of **generalisation** and those of **randomisation.** Randomisation methods include, for example, **stochastic superposition**, **swapping values** in a data set and **differential privacy**. Generalisation methods include **aggregation and k-anonymity**, **l-diversity** and **t-closeness,** as well as working with **synthetic data**. Synthetic data is data generated by computational methods without revealing the identity of a data subject. Which procedures are applicable to which data and which risks exist with regard to the attacker model is illustrated in the Practice Guide via an **evaluation**

matrix. There are technical methods of anonymisation that are **unsuitable from** the outset. This refers in particular to those methods that are based on **hash functions.**

2.7 Data transfer scenarios

In many cases, anonymised data is passed on to **third parties**. Likewise, service providers have an interest in using anonymised data for their own purposes. A recipient of such data will have to check whether the data is anonymous for him, taking into account the means available to him and the likelihood of their use. Contractual prohibitions on re-identification are not a criterion in the context of this objective review to exclude such re-identification per se. Anonymisation as processing of personal data, when involving third parties, may result in multiple **controllers being jointly responsible for such processing**. In such a case, the roles and responsibilities in the context of anonymisation must be clearly described in an agreement. **Delegating anonymisation** to a processor makes anonymisation vulnerable, as the controller can issue instructions to the service provider at any time in order to achieve disclosure of the technique used. The consequence would be that another body would have knowledge of the anonymisation technique, which an attacker could exploit.

If a service provider anonymises personal data of its client **for its own purposes,** it becomes a data controller. The consequence is that the controller needs a legal basis for the transfer and the service provider needs a legal basis for the processing of personal data at the same time. Moreover, the change of purpose must be **compatible** with the original purpose.

2.8 Four application classes

Practical scenarios of anonymisation of personal data can be divided into **four application classes.**

2.8.1 Application class 1: Anonymisation as deletion

Anonymisation as deletion is about removing a reference to a person in a data set in order to be able to continue to use data or properties in a data set. The Practice Guide gives three examples: The retention of key data from applications, quality analysis in the area of customer support and the creation of website statistics.

2.8.2 Application class 2: Disclosure of anonymised data

When **passing on anonymised data,** as a further application class, scenarios of "salary benchmarking", the passing on of sales figures according to product category as well as the matching of leaked access data are described, as is the analysis of a vehicle's fuel consumption.

2.8.3 Application class 3: Anonymisation when training algorithms

The third application class is dedicated to new technologies in the form of the **anonymisation of training data**. From a data protection perspective, the training of algorithms is problematic from several perspectives. In principle, the precise training of a model requires a broad database in order to provide sufficient information. Otherwise, there is a risk of creating a model that is too coarse.

In the context of data protection or the protection of company secrets, however, a consolidation of large amounts of data is often problematic, as these become known in the company that offers the creation of artificial intelligence models and thus end up in foreign hands. This is where **federated learning** comes in, for example, which can be applied in the context of training algorithms. In principle, federated **learning** means that the data can be used for training and also remain with the respective data owner. The service provider first creates an initial model, which it passes on to its partners. Using their respective data, each partner now tests the received model and informs the service provider how the parameters should be changed to improve the model. From all the feedback received, the service provider now calculates an overall update and applies it to the previous model, which is now distributed again.

In order to solve the problem of revealing personal data by observing changes in the evaluation of data sets in the area of federated learning, **differential privacy is an option, i.e.,** the avoidance of transmitting identifying characteristics about a person by selectively transmitting information from a database. The Practice Guide names synthetic data as another example of the application class of training algorithms. A good synthesis model is of elementary importance here, which may itself have to be trained.

2.8.4 Application class 4: Software testing

The fourth class of use is **software testing**, a process frequently encountered in practice. When generating test data that takes into account the properties of real personal data, it is important to ensure that the test data does not replicate the real data in such a way that it is possible to re-identify the data subject by using the test data. System migration and specific functionality tests, e.g. of user authorisations, are also examples of applications for anonymisation. An anonymisation solution that enables the testing of software functionality includes the provision of a test system that does not have access to real personal data.

In addition to determining a legal basis for anonymisation, data controllers must also comply with the **other legal requirements of the GDPR. In** view of existing accountability obligations, this includes the **documentation of** anonymisation in a separate concept or as part of the description of technical-organisational measures within the framework of the record of processing activities (RoP).

2.9 Special case: data protection impact assessment

Anonymisation does not per se pose a high risk to the rights and freedoms of natural persons. Other facts must be added which, taken as a whole, lead to a high risk. If there is a high risk, the performance of a data protection impact assessment is mandatory.

2.10 Transparency requirements

In order to maintain **transparency vis-à-vis data** subjects, **information** about planned anonymisation must be **provided** at the time of data collection or notification. The processing procedure of passing on already anonymised data to third parties, on the other hand, is no longer subject to the GDPR. Therefore, information about specific recipients or categories of recipients is no longer required. Data sharing on the basis of the EU law is thus also possible. If data has been collected without prior data protection information about the intended anonymisation pursuant to Art. 13 of the GDPR, e.g. in the case of older data files, Art. 6 (4) (e) allows for the following
GDPR allows for further processing if appropriate safeguards are in place. According to the wording of the GDPR, these also include encryption or pseudonymisation. If pseudonymisation can already constitute a suitable guarantee for use for a different purpose, anonymisation is a far more effective guarantee. Whether information about subsequent anonymisation must still be provided is assessed in accordance with Article 14 (5) (b) of the GDPR. This regulation is based on a **disproportionate effort.** Statistical purposes are cited as an example. With regard to the right of personality and data protection, the data subject is no longer at risk after anonymisation. Subsequent information about anonymisation does not create any added value for the data subject in terms of data protection law, but only generates a disproportionate effort.

2.11 Inspection obligations

Anyone who has anonymised data or uses anonymised data is obliged to **continuously check** that anonymisation is maintained. In doing so, it must be checked whether the personal reference can be restored. The implementation and the result of the check should be documented. Here, too, considerations of the attacker model must be included.

The above explanations show that data controllers and processors need clear descriptions in order to carry out anonymisation of personal data in practice. The Practice Guide therefore describes a procedural model for anonymisation, which at the same time represents the starting point for separate basic rules for anonymisation. These were also published in the course of this Practice Guide.

2.12 Procedure model for anonymisation

The process of an anonymisation procedure can be illustrated in the following procedure model.

Number	Measure	Chapter in the Guide
1	**Identify the legal basis for anonymisation (e.g., in the GDPR, in the BDSG or in a sector-specific law).**	4.2
2	**Ensure that the information obligations pursuant to Art. 13 and 14 of the GDPR are implemented.**	9.3
3	**Selection and determination of the appropriate anonymisation procedure**	
3.1	Type and risk class of personal data to be anonymised	6.1 and 8
3.2	Intended purposes of processing	6.1 and 8
3.3	Context of anonymisation	6.1 and 8
3.4	Expected number of records	6.1 and 8
3.5	Identify the statistical properties in the datasets that are needed and which characteristics are relevant for these properties	6.3 and 8
3.6	Determination of the appropriate anonymisation procedure and its timing	9.1, 9.2 and 9.4
4	**Conducting the anonymisation**	
4.1	Removal of all direct identifiers (e.g. name, address, contact details, credit card number).	3.2 and 6.1
4.2	Removal of all unnecessary indirect identifiers (e.g. gender, physical appearance characteristics, age, postcode).	3.2 and 6.1
4.3	Carrying out one or more procedures of the • Randomisation • Generalisation or • Carrying out a procedure with synthetic data.	6.36.3.1
5	**Analysis of whether and which risks exist for the restoration of the personal relationship**	6.1 and 6.2
6	**If risks exist, application of further procedures for anonymisation**	6.3.1
7	**Go through steps 4.1 to 4.3 until no more risks are apparent.**	6.3.1
8	**Check whether the required statistical Properties have been preserved**	6.3
9	**Document test and result**	9.1
10	**Use or share anonymised data**	

11	**Regularly repeat the test according to step 5, apply steps 6 and 7 if necessary and document according to step 9.**	

3. Terms and delimitations

3.1 What is meant by "data“

The term "data" means, among other things, "electronically stored characters, data, information".[3] The content of the stored characters, data and information is not important. The restriction to "stored" information in the Duden's definition is unfortunate. In the practical use of the term, **signs, indications and information are** also referred to as data if they are processed electronically but are not stored - not only temporarily.[4]

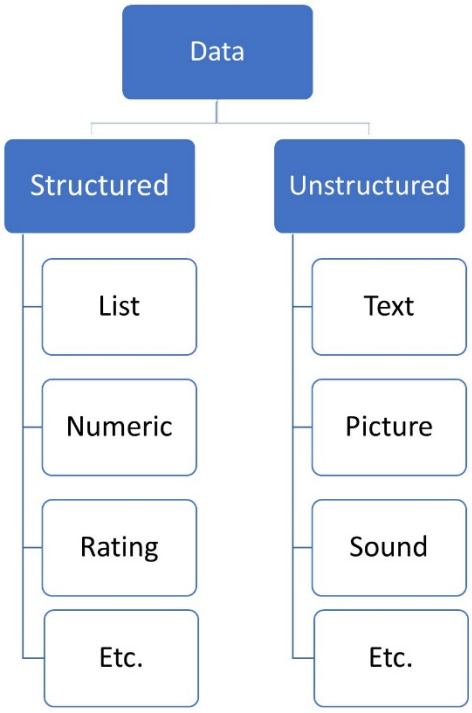


Figure 1: Types of data

Data exists in **structured** or **unstructured form** (Figure 1). Structured data regularly have the form "Key = Value" (e.g., "First name = Anne"). The key indicates what meaning the date should have. In the example, it would be "first name". The concrete content - here "Anne" is in the value part. Structured data can be understood as a key-value pair.

[3] Duden Online (2022): Data. URL: https://www.duden.de/rechtschreibung/Daten (last accessed 28.11.2022).

[4] See also Art. 2 (1) Data Governance Act (DGA).

Unstructured data is all data that does not correspond to a key-value pair representation. This includes, for example, comment fields in a customer relationship management system (CRM). Typically, any text can be entered in a comment field (e.g. "customer is nice" and "please ensure completeness when creating customer"). What information is contained cannot be predicted from the designation "remark". Other examples of unstructured data are e-mail, letters, pictures or sound recordings.

In practice, **mixed forms of** structured and unstructured data occur regularly. An e-mail, for example, contains the to-line, the subject and the sender as structured data. The so-called "email body", i.e., the text containing the actual message, is usually unstructured.

Especially in the case of structured data, it is important - with a view to the procedures for anonymisation - to distinguish which **value range** the data have. Frequently occurring value ranges are:

- **List:** Only values from a defined list of possible values occur (e.g., "yes" and "no" or "department management", "clerical", "management"). The order of the values on the list is irrelevant. If a list contains only two values, it is also called a Boolean value range.
- **Numeric:** Whole numbers or numbers with commas occur. One also speaks of quantitative data (e.g., height in cm or salary in euros and cents).
- **Rating:** The rating is a list of possible values. The values represent qualitative information (e.g., "1 - I like", "2 - undecided", "3 - I don't like"). Qualitative data cannot be added up or the like. 1 - I like" plus "2 - undecided" does not become "3 - I don't like".

Structured data is regularly stored in table form. Each row contains the details of a data record that represents, for example, a person. The individual columns contain data such as name, age, last purchase. These data in the columns are also called characteristics. Such a row is also called a **"data record".**

3.2 Personal data

Personal data is the counterpart to anonymous data. According to the legal definition from Art. 4 (1) GDPR, personal data are

- all information
- about
- an identified or identifiable natural person.

3.2.1 All information ...

On the one hand, there is **objective information** about a person such as name, address or date of birth. On the other hand, there is **subjective information** such as opinions, statements or assessments. The type and form of the information (e.g., alphabetical, numerical, as a photo or sound recording) are irrelevant from the perspective of the GDPR, as long as a meaningful content can be derived from it.

3.2.2 ... about ...

The information must be about a natural person. The existence of this requirement serves to exclude **factual information** from the GDPR. Factual data exists when information does not refer to a person but to a thing ("The red Flora cultural centre is located at Schulterblatt 71").

> **Note:** The subject date may not refer to a natural person, nor may the date be used to draw conclusions about a natural person.

The value of a property is also a material datum if it is used, for example, to describe the development of property prices in a certain region. However, if this property value is used to calculate the tax rate of a natural person, the tangible datum refers to a natural person. That means, an initial factual date may turn out to be information about a natural person at a later date or in a different context.

3.2.3 ... an identified or identifiable natural person

The information must relate to a **naturally identified or identifiable natural person.** A person is identified if he or she can be determined directly from the available information and can be distinguished from other persons (so-called **singling out**). The singling out of a person does not have to be based on a single piece of information. Existing information can also be combined to identify a person.

> **Example:** A customer list contains the first name, last name and place of residence of natural persons. Two entries have identical first and last names. By combining the place of residence with the first and last name, the person concerned can be identified directly.

With regard to the identifiability of a person, Recital 26 para. 3 of the GDPR states that *"account shall be taken of **any means reasonably likely to be** used, directly or indirectly, by the **controller** or by any **other person to** identify a natural person, such as singling out"*.

A person is **identifiable** if the existing information does not in itself allow for a unique identification, but a data subject can be identified by means of further processing steps and additional information or their combination. This can be done directly via the name or indirectly via a telephone number or a registration number. It is also possible to identify a person by delimiting a group to which the person belongs.

Example: A list with information on age and gross salary is available on the intranet of a company with 35 employees. When assessing the personal reference, it should be taken into account that employees often know the approximate age of a colleague and their position in the company. This means that with the knowledge of the age and the position, the respective salary can be determined.

The attribution to an identified or identifiable natural person does not have to be accurate in order to be considered personal. An **incorrect assignment** can also constitute information about a natural person. Furthermore, it is sufficient if a reference to a person can be made with a certain degree of **probability** (see section on this) or only for part of the data. 6.1) or only for a part of the data records. As soon as a reference to a person can be established in a case, all data records are to be considered personal.

Note: Whether information is about a natural person must be assessed on a case-by-case basis in relation to each piece of information subject to processing, taking into account the facts and surrounding circumstances.
For information to be classified as personal data, it is not necessary that the information in itself enables the identification of the data subject. It is also not decisive whether the name of a person is known.

3.3 Pseudonymisation

3.3.1 Definition and legal classification

Art. 4 (5) GDPR defines pseudonymisation[5] as

> *"the processing of personal data in such a way 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 which** ensure that the personal data are not attributed to an identified or identifiable natural person".*

The legal definition contains information on the legal classification of pseudonymised data as well as on the requirements for the process of pseudonymisation. It should be clarified that pseudonymisation is not a state, but rather a process that requires the **conversion of personal plaintext data into pseudonyms.**

[5] On pseudonymisation Schwartmann/Weiß (eds.), Whitepaper zur Pseudonymisierung der Fokusgruppe Datenschutz der Plattform Sicherheit, Schutz und Vertrauen für Gesellschaft und Wirtschaft im Rahmen des Digital-Gipfels 2017 as well as Schwartmann/Weiß, Anforderungen an den datenschutzkonformen Einsatz von Pseudonymisierungslösungen - ein Arbeitspapier der Fokusgruppe Datenschutz der Plattform Sicherheit, Schutz und Vertrauen für Gesellschaft und Wirtschaft im Rahmen des Digital-Gipfels 2018. See also in detail Schwartmann/Mühlenbeck in Schwartmann/Jaspers/Thüsing/Kugelmann, Heidelberger Kommentar DS-GVO/BDSG, Art. 4 DS-GVO Rn. 79 et seq.

If pseudonymised data is processed, it is - in contrast to anonymised data - still **personal data** which is protected by the GDPR. Thus, the processing of pseudonymised data requires a legal basis, in addition to the other requirements for the processing of personal data.

3.3.2 Requirements

3.3.2.1 No allocation of the data to a specific person without the use of additional information

Pseudonymised data must not be able to be assigned to a person without the addition of further information. That means, the pseudonymised data must not be information about an identified or identifiable natural person. In this respect, the general demarcation criteria between personal and non-personal and thus anonymous data apply.

Note: For pseudonymisation within the meaning of the GDPR, it is not sufficient if, for example, a reduced data set is passed on to a recipient, but the information from this data set can be assigned to a natural person.

3.3.3 Separate storage of the additional information

Pseudonymised data and existing additional information that enables re-identification of a data subject must be **processed separately**. If personal plaintext data is pseudonymised, for example, by means of cryptographic procedures, the responsible body must ensure that the cryptographic key is stored separately in order to re-establish the personal reference. Such separation can take place on a **logical level** (e.g., through authorisation concepts) but also on a **physical level** (e.g., by means of dedicated data processing systems) or on an **organisational level** (e.g., via a data trustee). The wording of the law does not require the data to be divided among several data controllers.

One-way functions such as hash functions (see chapter. 6.5.1) do not allow any "back calculation", i.e., the recovery of the plaintext data. However, a personal reference can be restored relatively easily by applying the one-way function to (presumed) plaintext data. If the result is the same as for the pseudonymous data set at hand, the pseudonymisation has been successfully removed. Therefore, only one-way functions should be used whose repeated application to the same data leads to different results (e.g., hash functions with **"Salt"** or **"Pepper"**). The settings to be made, if any, must be kept separately from the allocation lists.

3.3.3.1 Ensure technical and organisational measures for non-assignment

The separate storage of pseudonymised data and the additional information must be accompanied by **technical and organisational measures** (e.g., via an **authorisation concept** with different technical roles for access to the pseudonymised data or the additional information).

3.3.4 Use cases

Typical use cases of pseudonymisation can be found in the following areas:

- **Research**
  Example clinical study: In a clinical study, blood values of dialysis patients are examined. The identity data are pseudonymised in advance. If limit values are exceeded in a blood count, the person concerned can be contacted with the help of the pseudonymisation point and asked to undergo a specialist examination.
- **Analytics**
  Example user behaviour: A streaming provider evaluates the user behaviour of its customers. To do this, it pseudonymises the unique device identifiers in order to determine individual user behaviour.
- **Advertising**
  Example data matching: A company pseudonymises its customer data in order to match the pseudonyms generated from it with those of a social network. For this purpose, both entities use the same pseudonymisation procedure. The social network can now play out advertising to the company's existing customers without transmitting plain text data from the customer database.

3.4 Anonymisation

The GDPR does **not** contain **a legal definition** for anonymisation.[6] However, it follows by implication from the definition of "personal data" in Art. 4(1) GDPR and from Recital 26 pp. 3-5:

> *"[3] To determine whether a natural person is identifiable, account should be taken of **all the means reasonably likely to be used** by the controller or by any other person to identify the **natural person, directly or indirectly,** such as singling out. [4]In determining whether means are generally likely to be used to identify the natural person, all objective factors, such as the cost of identification and the time required for it, should be taken into account, taking into account the technology and technological developments available at the time of the processing. [5]The*

[6] Cf. however the definition in Sect. 4 LDG NRW, which is problematic under European law. See Schwartmann/Mühlenbeck, in Schwartmann/Pabst Landesdatenschutzgesetz Nordrhein-Westfalen, Sect. 4 para. 20 ff.

principles of data protection should therefore not apply to anonymous information, that is to say, information which does not relate to an identified or identifiable natural person, or personal data which has been rendered anonymous in such a way that the data subject cannot be identified or can no longer be identified. "

From the perspective of the GDPR, anonymisation is a **technical procedure** that is applied to personal data so that natural persons cannot be identified or can no longer be identified. The legal definition of personal data leaves open whether identifiability must be excluded for everyone or whether it depends, for example, on the respective controller. It is also not clear whether the state of anonymisation must exist for all time. In principle, various forms of anonymisation are conceivable.

3.4.1.1 Absolute anonymisation

If the reference to a person is practically impossible for anyone, this is called **absolute anonymisation**. In the case of absolute anonymisation, neither the person responsible nor third parties are able to re-identify the person concerned. It is technically, practically and factually possible for everyone, i.e. neither with the greatest possible effort nor by using any technical means. Absolute anonymisation is **the strongest form** of anonymisation. Achieving such a state is a great challenge. After all, all available means must be taken into account. This includes available data sources in the course of advancing digitalisation as well as increasing computing power.

The number of inhabitants in Germany is an example of absolute anonymisation. It is not possible to deduce from the number of inhabitants whether a particular person belongs or not.

3.4.1.2 Factual anonymisation

Factual or relative anonymisation is characterised by the fact that the re-identifiability of the data subject is not **completely excluded**. However, re-identification of the data subject is ruled out due to the **disproportionate effort involved,** taking into account the criteria mentioned in the GDPR as well as other criteria (see chapter. 6.1). In this case, the data is de facto anonymous for the controller or third party.

3.4.1.3 Case law on anonymisation

The European Court of Justice (ECJ) took a position on the controversial question of the reference to persons in its case law on **"Breyer"**[7] . In doing so, it essentially stated the following: It is not necessary that all information necessary for identification is in the hands of a single controller. Rather, it is sufficient if the controller has the data subject

[7] European Court of Justice, judgment of 9 October 2016, C-582/14.

identified via a third party. The limit of identifiability lies in the impossibility, disproportionality or violation of the law. The ECJ has thus essentially adopted a **relative understanding of the term, but has** allowed considerable exceptions. The challenge lies above all in the fact that ultimately, within the framework of the case law, the decisive question with regard to the reference to persons remains open as to when an impossibility or disproportionality of re-identification is to be assumed. [8]

3.4.1.4 Understanding of the terms of the GDPR

In view of the proportionality and likelihood considerations contained in the GDPR (see recital 26, p. 3 and 4), it is obvious that **de facto anonymisation** should also be considered in line with the legal requirements. This means that the use of the means and its probability is essentially assessed from the point of view of the person responsible. Any **additional knowledge of** a third party is not to be attributed to him per se. However, imputation is required if it is additional knowledge to which the responsible person **"could reasonably turn"**. This presupposes the knowledge of the responsible person about the third party and the knowledge and means available there.

Whether the use of **illegal means** should be included in the examination of the probability of identification is controversially discussed. Complete disregard does not appear to be appropriate, after all, an attacker (for the attacker model, see chapter 6.1) will not be able to identify himself. chapter 6.2) with the aim of re-identification will not be deterred from acting illegally. The overall assessment of the responsible person will be about how likely and how easy it is to use such illegal means.

If the anonymised data is to be **passed on** to recipients outside the controller **or access is to be granted to them**, it must also be checked whether the data is also anonymous from the perspective of these recipients. The knowledge and means of the recipients must be taken into account here. The recipients also include all persons or bodies who can legally obtain the anonymised data, e.g. through a right of access or inspection.

3.5 Anonymisation vs. pseudonymisation

The following illustration again clarifies the difference between anonymisation and pseudonymisation.

Pseudonymisation	Anonymisation
The person concerned can be identified again by obtaining additional information.	Affected person cannot or can only with disproportionate effort be identified again.
Processing is reversible.	Processing is irreversible.
The additional information must be kept separately.	

[8] For more details, see Schwartmann/Mühlenbeck, RDV 2022, 264.

The scope of application of the GDPR is opened for the entire processing activity.	Scope of the GDPR is left after successful anonymisation.

The smooth transition between an anonymous and a pseudonymous date is to be illustrated by the following example:

Example: Stadium operator T sells prepaid cards that enable cashless payment. A randomly generated card number is stored on the cards.

If the cards are purchased and topped up in the stadium without stating an identity (e.g. via payment with cash), they enable anonymous payment. The card number is therefore an anonymous datum.

However, if customers can additionally load the cards online, which requires registration with personal data and a link to the prepaid card, the identity of the customers can be determined by means of the additional information. For the stadium operator, the card number is now not to be classified as an anonymous date, but rather as a pseudonym.

3.6 Artificial intelligence

A computer programme, also called an **algorithm,** is - to put it simply - a set of rules in the style of "if x is true, then do y" and "do the whole thing z times". In this respect, a computer programme is no different from a cooking recipe or a click instruction. The rules built into the computer program reflect the developers' understanding of the subject matter processed in the computer program. This understanding may - but need not - be scientifically sound, correct or even ethically desirable.

Artificial intelligence" (AI) is also nothing other than such a computer programme. It has nothing - absolutely nothing - to do with "intelligence" in the human sense. Before the term "artificial intelligence" became a buzzword to suggest superior and modern technology, it merely referred to a subfield of computer science. This subfield deals with very different **classes of algorithms**. That is why there is no such thing as "the" AI, but rather a multitude of different types of algorithms and approaches, each of which functions differently and has different areas of application.

At its core, it is not a new or modern technology. The sub-field of "artificial intelligence" emerged in 1956[9] . In the past 66 years, the methods and algorithms have been constantly developed and made more powerful.

[9] McCarthy et al (1955): A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Grant proposal, August 1955, p. 1. URL: https://web.archive.org/web/20080930164306/http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html, last accessed 28.11.2022.

One of the most dominant classes of algorithms in public perception today is **pattern recognition by means of machine learning** (hereinafter referred to as "pattern recognition"). In some cases, AI is equated with pattern recognition in marketing, legislation, (legal) literature and in public.

Pattern recognition algorithms are characterised by the fact that neither the patterns to be recognised nor the features by which the patterns can be recognised need to be known in the programming phase of the algorithm. The algorithms for pattern recognition contain non-specific rules that are only adapted in a **"training phase"** so that with a certain probability the pattern considered "correct" by the developers is recognised.

Simplified, the training proceeds as follows:

1. The algorithm is presented with various data, including images, as input.
2. The algorithm "dices" the solution.
3. If the solution is correct, the algorithm receives a "reward" and if the solution is wrong, it is punished.
4. Adjust how you roll the dice
5. Start over with the next data set.

After sometimes millions of training runs, the algorithm recognises the training patterns with a **satisfactory probability**. Which patterns it can recognise is determined by the training data and its programming, i.e., it only ever recognises certain patterns defined by the manufacturer of the product in which the algorithm works. Training is usually finished at this point, i.e., no more "learning" takes place in productive use. Exceptions to this rule exist depending on the intended use.

Recognition does not mean that the algorithm understands the pattern semantically the way a human does. From the algorithm's point of view, it replaces the pattern with a word.

The rough process can be presented as follows using the example of image recognition:

1. Identify (presumably) related points in the picture (e.g., the same colour).
2. Consider related points as "object x
3. Search list of known objects for match with object x
4. Output the word y that is assigned to the object x in the object list

From the point of view of a pattern recognition algorithm, an image looks like a cloud of coloured dots. It does not "see" lines or geometric figures like a human being. A fortiori, it does not understand the meaning of the image.

The beauty of an algorithm is that it always provides an answer. Unlike humans, "don't know" is not part of its vocabulary. Therefore, the human observer easily gets the impression that the answer of a pattern recognition algorithm is "correct". This impression is deceptive, since "recognition" is a "guess" according to human understanding. The answers are not always correct. In other words, the recognised patterns are at best **more often right than wrong**. Those who use such algorithms must therefore reckon

with incorrect results and consider how to deal with them. Depending on how they are used, recognition errors lead to deaths. [10]

One measures the quality of the trained pattern recognition algorithm by means of the measures

- **Proportion of false positives:** A picture shows a horse, but the algorithm recognises a duck.
- **Proportion of false negatives:** An image shows a swarm, but the algorithm recognises a horse.

Even with correct answers from a pattern recognition algorithm, this does not mean that the algorithm pays attention to the same features as a human.

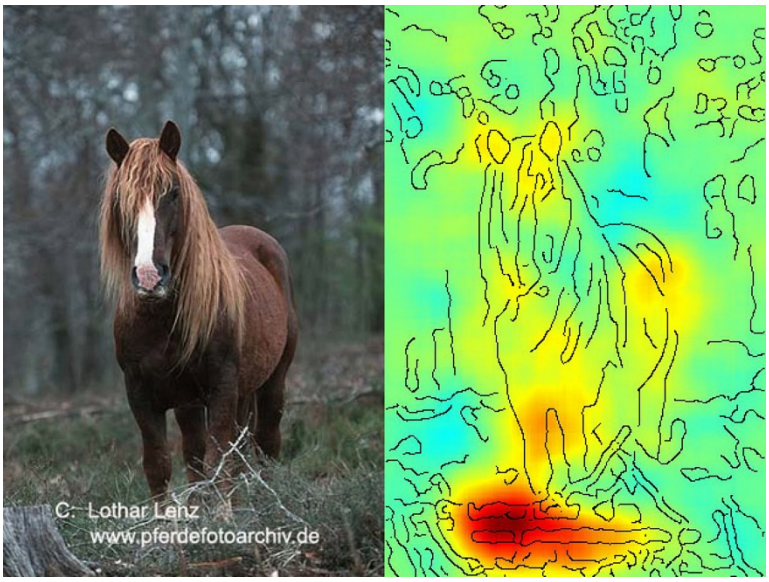


Figure 2: The image on the right shows the analysed image areas of the image on the left in red.[11]

The right image in Figure 2 visualises the sections of the left image that are used to recognise a horse. The copyright notice is used because training images with horses are distinguished from other images by the copyright notice.[12] If the algorithm were now shown a picture of a house with the same copyright notice, the algorithm would recognise the house as a horse.

Because the rules for recognising patterns are not explicitly written into the **source code** of the algorithm by the software developer but are only "formed" during the training

[10] Cf. e.g. Der Standard (2022): "Full Self-Driving": Another fatal Tesla accident with activated Autopilot. 03.08.2022, URL: https://www.derstandard.de/story/2000137996067/full-self-driving-erneut-toedlicher-tesla-unfall-mit-aktiviertem-autopilot (last accessed 28.11.2022).

[11] Source: Heise Online (2020): How AI decisions can be verified. URL: https://heise.de/-4665982 (last accessed 28.11.2022).

[12] Heise Online (2020): How AI decisions can be reviewed. URL: https://heise.de/-4665982 (last accessed 28.11.2022); Further reading: Christopher J. Anders, Talmaj Marinc et al. (2019): Analyzing ImageNet with Spectral Relevance Analysis: Towards ImageNet un-Hans'ed arXiv:1912.11425, 2019.

phase, it is not possible to understand how a result of the algorithm comes about without tools. In algorithms that are developed without machine learning, the rules are explicit and, in principle, easy for a human to read like a book in the source code. Given enough time and knowledge, any human being can understand how the algorithm works by reading the source code.

Such a possibility does not exist for algorithms developed by means of machine learning. Additional technical tools are needed to make the functioning transparent. For this reason, errors are often only found in productive use, if they are detected at all. Sometimes with fatal consequences for the people involved.

4. Legal environment

4.1 Anonymisation in the light of the European Data Strategy

Data processing in the EU will be embedded in a wider context in the future. The central legal acts are the already adopted **Data Governance Act (DGA)** and the emerging **Data Act (DA)**. Under the former, public bodies will be able to make data available for further use. The latter addresses the economy. Data that is currently in the hands of manufacturers of **IoT devices** should also be made commercially usable for users and other companies. To this end, users will be enabled to share their data from such devices. In this way, the legislator wants to create incentives for innovative business ideas in the right place. In addition, the **Digital Markets Act (DMA)** imposes obligations on the gatekeepers that dominate the digital economy within the EU in order to create fair competition in the internal market. Users are to be given more data sovereignty. The **Digital Services Act (DSA)**, in turn, claims no less than to safeguard democracy. In particular, it obliges the major online platforms to fight hate, fake news and crime on the internet. The corporations must establish procedures that mitigate the risks of their business model. The EU also attaches particular importance to the **regulation on artificial intelligence (AI)** - it is to be adopted in 2023 and make the EU the global trendsetter in the fair use of this key technology. Finally, the EU's activities also extend to internet security. The draft revision of the **Network and Information Security Directive (NIS)** provides for a significant expansion of the original scope and now also covers smaller companies. New obligations are also defined here, as in the **Radio Equipment Directive (RED)**. This addresses the dangers of networked end devices (IoT) and must be implemented by 2024. A **Cyber Resilience Act (CRA) is** also planned, which is to guarantee the technical stability of products used in the network.

The data-sharing mechanisms provided for in the new legal acts mostly offer the possibility to implement safeguards for the benefit of data subjects. These also include the **pseudonymisation** and **anonymisation of** personal data. The anonymisation of data will therefore gain significantly in importance with the new data acts. The GDPR remains the **standard** when it comes to processing personal data, which must be complied with in addition to specific requirements of the new acts. Laws in the EU that concern the protection of personal data are not affected by the GDPR. Thus, the statements of the

GDPR continue to apply to the important distinction between pseudonymisation and anonymisation of personal data.

4.2 Anonymisation as data processing within the meaning of the GDPR

Unlike pseudonymised data, the GDPR does not apply to anonymous data. However, anonymisation is characterised by the use of a technique to transform personal data into anonymous data. The initially collected personal data must be processed in accordance with the GDPR, including the requirements of a **legal basis.** Whether a legal basis is also required for anonymisation itself must be assessed on the basis of the scope of application of the GDPR. It applies when personal data are processed in whole or in part by automated means or when non-automated processing takes place and the data are stored in a file system. If one consults the definition of **processing** in Art. 4(2) GDPR, it is obvious that the transfer of the status "personal" to "anonymous" is regularly regarded as processing in the form of a change of personal data. This means that the legal requirements of the GDPR must be observed for the anonymisation process, including the existence of a legal basis (for the other legal requirements, cf. chapter 9).

The **table** provides an overview of the legal bases that are usually relevant. Which legal basis allows a specific anonymisation must be assessed on a case-by-case basis.

Data type	Purpose "anonymisation" communicated during collection	Purpose	Common legal basis in the GDPR
-/-	-/-	Deletion	Legal obligation (Art. 6 para. 1 lit. c) in conjunction with 17 para. 1 lit. a))
"normal"	yes	Disclosure to third parties	• Consent (Art. 6 para. 1 lit. a)) • Weighing up interests (Art. 6 para. 1 lit. f) - only for non-public bodies)
"normal"	no	Disclosure to third parties	Check whether new purpose is compatible with original purpose of processing (Art. 6(4))
Special categories (Art. 9 (1) GDPR)	yes	Disclosure to third parties	• Consent (Art. 9 para. 1 lit. a)) • Legal obligation under labour or social law (Art. 9 para. 1 lit. b))
Special categories (Art. 9 (1) GDPR)	no	Disclosure to third parties	Not permitted
Criminal convictions & offences (Art. 10)	-/-	Disclosure to third parties	Basically inadmissible

Table 1: Overview of commonly applicable legal bases [13]

[13] Lepperhoff, N. (2022): Anonymisation of personal data - Part 2: Practical examples and boundary conditions. In: Lohn und Gehalt 07/2022.

5. Functions of anonymisation

Anonymisation prevents data from being assigned to specific persons. For this purpose, the process of anonymisation **removes, replaces, aggregates or falsifies** personal data or data that can be related to persons (see chapter 6.3.1). In the case of anonymised data, this prevents re-identification. For completely anonymised data, the requirements of the GDPR no longer apply (see chapter 3.4). The data controller is therefore no longer bound by data protection regulations on the permissibility of processing anonymised data. This means that further use for analysis or disclosure is generally permissible. The extent to which anonymisation is suitable for a data controller's desired purposes depends largely on the **technology** used and the desired **usability of data** (for selected classes of use, see chapter 8).

Note: A controller is not prevented from using anonymisation as a **technical/organisational measure to** mitigate the risk to the rights and freedoms of data subjects. Especially in situations where there is uncertainty about the risk of re-identification, anonymisation can still be useful. The scope of the GDPR would then not be left, but personal data processing can be enabled by this measure. Anonymisation is also relevant to the requirement of **data minimisation** pursuant to Article 5 (1) (c) of the GDPR.

6. Requirements for anonymisation

6.1 Legal

The GDPR does not contain any specific requirements as to when personal data have been sufficiently anonymised. The definition of personal data in Art. 4(1) of the GDPR and the explanations in Recital 26 of the GDPR (on personal data cf. the chapter on personal data) do, however, contain some indications and provide guidance on the requirements for anonymisation from a legal perspective. 3.2) do, however, contain some indications and give hints as to which requirements are to be placed on anonymisation from a legal point of view.

In view of Art. 4(1) of the GDPR, the data set resulting from anonymisation must not contain any information about an identified or identifiable person.

Whether the data generated from anonymisation relates to an identified or identifiable person is specified in Recital 26 pp. 3 and 4 of the GDPR:

> *"In order to determine whether a natural person is identifiable, account should be taken of* ***all the means reasonably likely to be used by*** *the* ***controller or by any other person to identify*** *the natural person,* ***directly*** *or* ***indirectly,*** *such as* ***singling out****. In determining whether* ***means are reasonably likely to be used to*** *identify the natural person, account should be taken of all objective factors, such as the cost of identification and the time required for it, taking into account*

the technology and technological developments available at the time of the processing."

The body carrying out anonymisation is first **required** by the recital to **verify** whether or not the operation involves personal data. In the context of this examination, **all means** must be taken into account that could **reasonably** be used either by the controller or a third party to identify the data subject. Such means may be, for example, information available to the controller or information that the controller can obtain. In this respect, he will also have to take into account the **linking** of so-called indirect identifiers, which may lead to the identification of a data subject. **Contextual information** or **raw data** can also play a role. Especially in view of the European Data Strategy, publicly accessible data rooms with personal, pseudonymised or anonymised data are increasingly to be expected. A variety of data sources can increase the **probability of re-identification.** It is irrelevant whether the data controller or a recipient of the data wants to identify the person or not. **Objective identifiability** is sufficient.

There are cases in which the probability of the use of means cannot be answered easily. Here, the GDPR names some **factors** that can be used as test criteria with regard to the respective means available at the controller or at the third party:

- Identification costs
- Time required
- Technologies available at the time of processing and their development
- Other objective factors

From a technical point of view, a controller will have to ask himself whether there are technologies available at the time of processing, i.e., anonymisation, which could favour re-identification. Accordingly, the anonymising body must also use a procedure according to the **state of the art**[14] . Since technologies do not remain static, their **developments** must also be included in the analysis. For example, if a responsible party deletes the key used to encrypt a data set, the data set should not be considered anonymous for all time. After all, technological progress (keyword: **quantum computing**) may lead to a reversal of encryption in the future.[15]

If the controller has the **legal possibility** to (re-)identify a person by exercising rights, this must be included in the probability test.

The **interest in re-identification** is regularly reflected in the value of the data to an attacker. This can also be included in a probability test. For example, the reversal of

[14] "The state of the art is the level of development of advanced processes, equipment and modes of operation which, according to the prevailing opinion of leading experts, makes it appear certain that the legally specified objective will be achieved. Processes, equipment and modes of operation or comparable processes, equipment and modes of operation must have proven themselves in practice or - if this is not yet the case - should, if possible, have been successfully tested in operation." (cf. Handbook of Legal Formalities of 22.09.2008, marginal no. 256).

[15] See EDPS and Agencia Española de Protección de Datos, 10 Misunderstandings related to Anonymisation, available at https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf (last accessed 28.11.2022).

anonymised credit card records is of great interest to an attacker, which must have implications for the strength of the anonymisation. Health data can also be of particular interest to an attacker, e.g., in order to determine the illness of a specific person (for the attacker model, see the chapter on the attacker model). chapter 6.2).

Whether anonymised data is passed on to **internal recipients** or such data is **published** should have an impact on the risk analysis of re-identification. In the case of publication, there is regularly a large **group of recipients** who may be able to re-identify a data subject by means of additional information. Therefore, when publishing anonymised data, strict standards should be applied to the said risk analysis.

New means may also have to be considered at a **later point in time,** the existence of which may have an influence on the question of sufficient anonymity. Anonymisation procedures must therefore be **reviewed** and **evaluated on an** ongoing basis (on review obligations cf. chapter 9.4).

From the perspective of the GDPR, in the absence of legal specification, various **anonymisation techniques** can be used. The decisive factor is that, after examining the

Note: Removing **direct identification features** (e.g. a person's name) is in many cases not yet sufficient to anonymise personal data. It is also possible to identify a person from other available information (e.g. gender, occupational group and year of birth). If the controller himself or a third party can re-identify or make identifiable data subjects without disproportionate effort, personal data are not sufficiently anonymised per se.

factors listed above, re-identification of data subjects is not practically feasible. This means that, if it requires a disproportionate effort in terms of time, costs and manpower, effective anonymisation can generally be assumed.

As long as the personal data set used for anonymisation is available to the controller, the data anonymised from it will also remain personal to the controller on a regular basis. This means that the GDPR and all other data protection laws also apply to the anonymised dataset if the anonymisation can be easily revoked by accessing the original dataset.

Note: From a legal point of view, a responsible person will have to examine the factors named in the law as well as all other circumstances in an **overall view** as to how likely it is that someone will make the effort of re-identification.

6.2 Attacker model

In order to assess whether data is anonymous, it is useful to take the perspective of an **"attacker" “”** who tries to re-establish a personal reference or derive statements about a concrete person from the data.

An **"attacker model"** thus describes a method for checking whether a data set is anonymous or personal.[16] From the perspective of an attacker, it is tested whether re-identification is possible. Only if such an attempt - carried out seriously - fails, can one speak of anonymous data.

The **knowledge** and **skills** assumed of the attacker depend on the **context of use** of the data. If the data is to be made public, it can be assumed - in view of the multitude of (criminal and state) actors - that the attacker has deeper expertise and a higher level of equipment than if the data is given to a specific recipient. An attacker can pursue various goals in order to re-identify persons in an anonymised data set. The more valuable the data can be for the attacker, the more expertise and resources are to be assumed. Not all theoretically conceivable or non-excludable technical possibilities or possibly existing knowledge are to be included, but those that are reasonably probable.[17]

In principle, different **types of attack** are conceivable:

- An attack based on **existing knowledge** about a person and known to exist in a data set.
- An attack that uses **public or other sources at its disposal to** re-identify the person.
- An attack that aims to re-identify **as many people as possible** from the dataset, even if this means falsely identifying people.

Note: It is advisable to consider not only **targeted** attacks in the attacker model, but also constellations in which re-identification by the attacker is actually **unintentional or** could occur by **chance.** Likewise, there may be attacks that aim to "interact" with the persons concerned without knowing who they are (e.g. by obtaining additional information about a person from the data set).

There are various **implementation options** for carrying out an attack with the aim of re-identification:

[16] Art 29 Group (2014): Opinion 5/2014 on anonymisation techniques. Adopted 10 April 2014. WP216, ICO (2021): draft anonymisation, pseudonymisation and privacy enhancing technologies guidance, Chapter 2, p. 14ff.

[17] ICO (2021): draft anonymisation, pseudonymisation and privacy enhancing technologies guidance, Chapter 2, p. 12.

- **Singling out**: Selected data records are singled out from a data set in order to identify a person. The person concerned can be sufficiently distinguished from other persons on the basis of the information available in the data set or accessible to the attacker. For example, in a salary list, the managing director may be the person with the highest salary.

  Example: An attacker infers the assignment of a person to such a characteristic due to insufficient diversity in a dataset, here regarding sensitive indirect identifiers (e.g. "person X has cancer"). Such an attack is favoured if there are not enough variations of sensitive characteristics in a group of persons (see also k-anonymity under chapter 6.3.1.2.1 and l-diversity under para.6.3.1.2.2).

- **Linkage of data records ("record linkage")**: The attacker attempts to link a record from the anonymised database to a person. With background knowledge about indirect identifiers, the attacker can link a small group of records or possibly a single record to a person. The link can also be made using statistical methods, for example. It is sufficient to establish a personal link that there is a probability that two data records belong to the same person.
- **Inference**: A (new) characteristic of a person is derived from the values of other characteristics available in the database. Statistical methods are used to search for relationships between characteristics. It is sufficient for a personal reference if the presumed relationship is significantly probable.

  Note: The attacker model must include all data sets to which the controller, recipient of the anonymised data and the attacker have or may have access.

6.3 Technical

Anonymisation methods **apply data transformations to the original data** to achieve the desired characteristics of the anonymous data set. That is, **identifiers** (both direct and indirect) are removed and/or transformed so that a dataset of information can no longer be attributed to a specific person.

Anonymisation procedures walk a fine line. On the one hand, the personal reference should be removed. On the other hand, the **statistical properties** of the data should not be changed. An anonymisation procedure necessarily changes the data in order to remove the reference to a person. The decisive question is whether the change also leads to an undesired change in the statistical properties.

Not every anonymisation procedure is suitable for every application. If an unsuitable procedure is used, the personal reference can remain or the personal reference is removed together with the required statistical properties. The result would be useless in the best case.

When selecting one or more anonymisation methods, the user should determine which statistical properties of their data should be preserved in any case. This also means determining and technically understanding the statistical properties contained in the data. Such understanding depends on the application area and the concrete data.

Often, **several anonymisation procedures** have to be carried out in succession. What these are and what the sequence might look like also depends on the individual case.

At this point, the anonymisation procedures are therefore presented in abstract form to give an impression of the area of application and the mode of operation. Whether ready-made products or open-source implementations are available on the market for concrete use cases would have to be checked in each individual case. Whether a product is suitable should be carefully examined. If no product is available, the only option is to implement it oneself using software technology. Furthermore, the overview represents a snapshot, as research on anonymisation procedures is ongoing.

Anonymisation procedures can be divided into two classes:
- Randomisation
- Generalisation

In the following, known procedures are presented that can anonymise existing personal data.[18] In the field of **machine learning,** other procedures are known that do not anonymise data, but reduce the depth of intrusion into the personal right in the training process or training result. These methods are presented in chapter 8.3 presents examples of these methods.

6.3.1 Selected procedures

6.3.1.1 Randomisation procedure

At its core, randomisation changes values randomly. This change leads to the removal of an association between different characteristics. Inference risks are reduced. The way the changes are made depends on the method chosen.

6.3.1.1.1 Stochastic superposition

Stochastic superimposition changes the values of individual characteristics in a data set. The prerequisite is that the values are **numerical, i.**e. quantitative. Furthermore, it is assumed that the original data is deleted after the procedure is applied so that the change cannot be traced.

[18] The presentation is mainly based on Art 29 Group (2014): Opinion 5/2014 on anonymisation techniques. Adopted on 10 April 2014 (WP 216).

The change is not made arbitrarily, but in such a way that the statistical distribution of the original values is not changed. If the original value occurs in 5 out of 100 cases, the changed value should also be included in 5 out of 100 cases. For the change to be irreversible, it must be unpredictable - random. For example, it is not sufficient to increase the height of a person by the fixed value of 5 cm, since the original values could be calculated by subtracting 5 cm. If instead of a fixed value a random value from the range -15 to +15 cm were chosen, the original values would not be calculable.

Critical for the success of the method is the choice of how the values are changed **("disturbance variables")**. If the choice of disturbance variables leads to exaggerated results, e.g., body heights of 250 cm for people, the changed characteristics can be determined on the one hand and the change - at least on the basis of assumptions - can be calculated out. For example, one can assume that the 250 cm belong to a very tall person. If one assumes that a very tall person is 210 cm tall, the disturbance size would be 40 cm.

If logical connections between features of a data set are removed by the stochastic overlay, an attacker can also use this knowledge to reconstruct the person reference.

The change leads to a **loss of information**. The amount of this loss can be calculated as the quotient of the maximum possible change - in this case 15 cm - and the maximum basic value - e.g., 210 cm. In the example, the information loss would be 15 cm/210 cm = 0.07 = 7 %. Whether the loss of information is sufficient for sufficient anonymisation must be assessed in each individual case.
The use of stochastic superimposition is often not sufficient for anonymisation, i.e., it must be supplemented by **other techniques.**

6.3.1.1.2 Exchange

The **swapping** leaves the values of the characteristics untouched. Rather, values are swapped between the data sets. This makes the method suitable for both **qualitative data** (rating, lists) and **quantitative data**. For example, the height is swapped between data set "154" and data set "357". The prerequisite is that the original data is deleted after use.

Swapping removes the **association** between **feature and record.**
If only one data field is swapped from a logical relationship or statistical correlation, the personal reference can be restored. To do this, it is only necessary to know which swapped characteristics are related to non-swapped characteristics. For example, the salary is swapped between "154" and "357". The position of the person is not swapped. Furthermore, a random swap does not necessarily cancel out strong **statistical relationships** between characteristics. Therefore, after the swap, it must be checked whether the person reference can be re-established by exploiting such relationships.

If the swapping is done separately for each characteristic, the (statistical) properties of the data set can be changed. Therefore, the characteristics whose (statistical) relationship is to be preserved must be swapped between the same data sets. For example,

the salary and the position are swapped between the data sets "154" and "357". Since the values of the data remain unchanged, the information content does not change.

Not every interchange automatically leads to anonymisation. Care must be taken to ensure that the characteristics that are the cause of the personal reference are swapped.

6.3.1.1.3 Differential Privacy

Differential Privacy[19] does not change the original data, so it does not have to be deleted after the procedure has been applied. The original data remains unchanged and personal. The prerequisite is that the data are **quantitative.**

The concept of differential privacy provides different users with a **limited view of the** original data set. The number of data records is restricted. For the displayed data sets, the values of the characteristics in the display are changed by means of mathematical functions. This means that the user sees different data than in the original dataset. The type of change and how it is done technically depends on the individual case.

The concept of differential privacy merely provides a mathematical framework and procedure to determine the change. The change is recalculated for each query.

It is true that the same data is always changed differently with each query. Nevertheless, the combination of multiple queries allows the mathematical change of the values to be determined and "calculated out". For this reason, it must be ensured that multiple queries of a body are prevented. When assessing whether the result is anonymous, access to the displayed data must also be taken into account.

6.3.1.2 Generalisation procedure

In a **generalisation,** values are "coarsened" in their order of magnitude. For example, a street name is replaced by the postcode. This "coarsening" is intended to make it more difficult to pick out people. The other risks of linkability and inference do not change in principle. In order to preserve the statistical properties as much as possible, the "coarsening" is done regularly.

6.3.1.2.1 Aggregation and k-anonymity

With **aggregation** and **k-anonymity,** data records are combined into groups. Data records in a group receive the same characteristic value. For example, salary data is replaced by intervals "20-30,000" and "30-40,000".

[19] See also Ostendorff, OpenRedact Anonymisation Guide Open Data and Data Protection, available at https://openredact.org/leitfaden-anonymisierung (last accessed on 28.11.2022).

If the values of the characteristic to be aggregated are not equally distributed, but some values occur significantly less frequently or more frequently than others, there is a risk that there will be groups containing only one data set. To prevent this, k-anonymity specifies that each group must contain at least "k". The parameter k describes the minimum size of a group. There is a trade-off between group size k and information content. The larger k is, the smaller is the information content. For example, the interval "1-100 million" would not be very meaningful as a salary indication.

Often, **several characteristics** can lead to an identification of the person. Therefore, all characteristics that could be suitable for an identification of the person ("indirect identifiers") are to be included in the group formation. A group then consists of several characteristics. Otherwise, the unaltered values of the indirect identifiers can be used to identify persons from a group. If, for example, in the context of a medical study, the plain names of the test persons are replaced by a character string, but the disease diagnosis, postcode and date of birth are also given, the date of birth can be replaced by the year of birth to create k-anonymity. If there are three persons in the data records whose entries match with regard to certain characteristics, e.g., the same year of birth or the same postcode, the k-anonymity is 3. A (clear assignment) of the disease diagnosis to a specific person is no longer possible.

In practice, the challenge is to take into account all indirect identifications, i.e., these must be determined beforehand. Special attention must also be paid to characteristics that have a high information content, e.g., because they have very rarely occurring values.

A k-value that is too small also calls anonymisation into question. When forming groups, care must also be taken that individual data records are not given too much weight. This risk exists especially with an uneven distribution of the values of characteristics.

6.3.1.2.2 l-Diversity and t-Closeness

The concept of **l-diversity** further develops the concept of k-anonymity. The k-anonymity determines the minimum group size k. How often an individual value occurs in the group is not specified.

> **Example:** Age and salary are stored for employees of a company. The concept of k-anonymity is applied to the salary. In the group "20-30,000 euros" are the persons A (30 years), B (35 years) and in the group "30-40,000 euros" are the persons C (65 years) and D (43 years). If an attacker combines further knowledge with the data set, he recognises that only one 65 year old works in the company. Thus, the attacker concludes that the 65-year-old earns 30-40,000 euros.

The personal reference could be established in the example because the value 65 years occurred once in the group. The l-diversity requires that in each of the "k" groups each characteristic has at least "l" different values. Instead of one 65-year-old, there should have been "l" 65-year-olds in the group. The prerequisite for applying l-diversity is that a **sufficient number of data sets** in the dataset have the **same values.**

If one takes as a third requirement that the distribution of values in a class should correspond to the distribution of these values in the original data, one speaks of **t-Closeness**. The prerequisite is that the dataset has a sufficiently large number of suitable values.

Taken together, the data records are to be divided into groups in such a way that there are at least "k" data records in a group, each value occurs at least "l" times and that each value in the group occurs as often as it occurs in the original data record considered over all data records.

6.3.1.3 Synthetic data

The data protection requirements to be observed when processing personal data may limit the extent and type of data processing. A possible alternative to the use of personal data for certain use cases is the use of **synthetic data**. In contrast to personal data, synthetic data are not collected on specific natural persons. Accordingly, they do not provide information about natural persons. Rather, synthetic data is **data generated** by a **calculation process**.

Synthesis models are used to generate synthetic data. These determine the essential properties of the calculation method used to generate the synthetic data. The known synthesis models include

- Models that randomly select data from a list of given examples. An example is the random selection of cities from a list of entries that cannot be assigned to specific data subjects.
- Models that generate synthetic data from the random concatenation of strings from an alphabet
- Models that generate data according to rules that are fixed by humans. An example is the rule to generate data in a certain format or the rule to select data from a list of female first names of French origin.
- Models that generate data according to rules determined from an AI. These can be correlations that the creator of the calculation procedure would not have recognised without the help of the AI.

Synthetic data consisting of several data fields can also be generated from a combination of the three types.

Synthetic data can be **modelled** on personal data in certain properties. These properties are simulated by the calculation procedure in such a way that the calculated synthetic data can be used instead of the personal data. Usability requires that the synthetic data resemble the real personal data. The degree of similarity is determined by the use case. At the same time, it must be noted that the synthetic data do not allow for re-identification of data subjects.

An example of the use of synthetic data is data that is to be used to test the functionality of newly developed software to be used with personal data. Instead of testing software

with real personal data, synthetic data is used. In order to obtain meaningful test results, these must be **sufficiently similar** to the real data. The similarity is achieved, among other things, by reproducing **statistical properties of** the personal data by the calculation procedure when generating the synthetic data. In doing so, it must be ensured that the synthetic data do not resemble the personal data in such a way that they can lead to the re-identification of data subjects. For this, the calculation procedure must be carefully designed. In addition to preserving certain properties of the underlying personal data, care must be taken when designing the computational procedure to ensure that the replication of the properties does not result in synthesised data that could lead to re-identification of data subjects. An example is the replication of a data set consisting of first and last names of female persons from the German language area. When word pairs (first name, surname) are transferred from the underlying real data set to the synthetic data set without any changes, rarely occurring name combinations in particular can be used to re-identify affected persons.

6.4 Evaluation matrix

Technology	Applicable to	Singling out	Linkability	Inference
Stochastic superposition	1	-	-	o
Exchange	1,2	o	-	-
Differential Privacy	1	+	- (multiple application)	- (multiple application)
Aggregation and k-anonymity	1,2	+	o	-
L-Diversity and t-Closeness	1,2	+	o	o
Synthetic data	1,2	+	+	+

Table 1: Overview of the effect of anonymisation techniques (1 = quantitative data, 2 = qualitative data, - still possible, o possible but difficult, + prevented)[20]

6.5 Differentiation from other procedures

6.5.1 Hash function

A **hash function** is a mathematical procedure that basically replaces a character string with a shorter character string. A name, such as "pattern affected person", becomes "AF341".

[20] The presentation is mainly based on Art 29 Group (2014): Opinion 5/2014 on anonymisation techniques. Adopted on 10 April 2014 (WP 216).

If an attacker knows which algorithm was used to calculate the hash values, he can calculate the hash values of the values he is interested in and compare his calculated hash value with the "anonymous" hash value. If both hash values match, he has probably determined the original date, in the example the name. It is not a safe determination, as "hashing" different original values can lead to the same hash value. To make this "back-calculation" more difficult, cryptographic hash functions can add further random values ("**salt"** and **"pepper"**). This would always generate a different hash value when "hashing" "pattern affected persons". If the **attacker** learns how to calculate Pepper and Salt, he may be able to re-identify them as well. In any case, the static properties of the hashed feature are destroyed.

Data cannot be made anonymous by means of hash functions.[21] The result of the hash process is pseudonymity at best.

7. Involvement of third parties or processors

7.1 Disclosure to third parties

It is possible that anonymised data is **passed on** to a **third party** and thus an independent controller. If the data are sufficiently anonymised prior to transfer, the data protection requirements of the GDPR in general (such as the permissibility of the transfer or its transparency vis-à-vis data subjects) do not have to be observed. However, the third party as recipient of the data will have to check whether the data is anonymous **for him**, taking into account the means available to him and the likelihood of their use. A **contractual prohibition** of re-identification cannot be an effective measure to exclude such per se. If the data transferred are not anonymous for him, the processing again falls within the **scope** of the GDPR.

Incidentally, there is no legal obligation to involve a third party, e.g., as a **trust agency,** in anonymisation.

7.2 Joint controllers

If two or more controllers **jointly decide** on the **purposes and means of** processing personal data and anonymisation forms part of the processing operations, the requirements of Art. 26 of the GDPR for the agreement between the parties involved must be observed. The **framework conditions for anonymisation** should be described there. If a controller carries out anonymisation within the framework of joint responsibility, he or she should be named in the agreement as the controller for the anonymisation procedure. If this person in charge grants access to anonymised data, a **rights and roles concept** must ensure that access to an original data set by the other persons in charge is excluded.

[21] Other view: BDI (2020): Anonymisation of personal data, p. 21.

Furthermore, the **contractual obligation of** the parties involved to **check whether a** natural person can be identified from an existing own data stock by comparison with an anonymised data set transmitted, for example, is a good idea. In the event that re-identifiability is detected, the consequences should be regulated in the agreement. For example, the revival of the requirements of the GDPR with regard to the processing of personal data.

7.3 Processor

The GDPR recognises two roles that companies can assume: **Controller** and Processor. For example, a service provider can carry out payroll accounting for employers. For this payroll on behalf of the employer, the service provider acts as a processor. However, for its own recruiting process, it is the responsible party. A processor acts as a vicarious agent for its client. Thus, a processor is a **service provider bound by instructions**. A controller determines the purposes or means of processing. He is the "master of the data".

7.3.1 The processor anonymises for the controller

If personal data is anonymised by a processor in accordance with Art. 4(1) of the GDPR (e.g., by an anonymisation tool hosted by the processor), the requirements of Art. 28 of the GDPR must be observed. After all, personal data is being processed here (see also chapter 4.2). This case must be distinguished from an anonymisation tool operated by the controller itself.

However, it is questionable whether commissioned processing is a **suitable means for** anonymisation in view of the fact that the service provider is **bound by instructions.** After all, the controller could theoretically oblige the service provider to **disclose the anonymisation technique used**. In addition, another body may have **knowledge of the anonymisation** technique used, which an attacker could exploit. If a controller wants to have personal data anonymised by a processor, **contractual provisions** must at least be made that prohibit disclosure of the anonymisation technique and its implementation steps. It is more advantageous to operate anonymisation software oneself.

7.3.2 The processor anonymises for its own purposes

7.3.2.1 Admissibility of the processing

What happens if the service provider bound by instructions wants to use data for its own purposes and anonymise them beforehand for this purpose?

Anonymisation is a data processing for which there must be a permissive element in the GDPR.

As soon as a processor uses the client's data for his own purposes - even if it is "only" through anonymisation - **he becomes the controller for this processing** (Art. 28 (10) GDPR). The service provider bound by instructions becomes the "master of the data".

The processing of the service provider for its own purposes means, with regard to the lawfulness of the processing, that

- the original controller has to check the **compatibility of the change of purpose** (Art. 6(4) GDPR) and
- the service provider requires its own **legal basis for the** data processing.

Whether the service provider has a **legal basis for** processing for its own purposes must be carefully examined. This also applies to the processing operation of anonymisation. After all, the service provider is closely **bound by contract and instructions with** regard to data processing.

In some cases, processors contractually grant themselves access and utilisation rights to the client's data in general. This granting of rights alone does not legitimise the associated data transfer. The granting of a right for self-interested data processing by the service provider must be **compatible** with the original legal relationship of the controller to the data subject. The consent of the data subject for the data processing and thus also for the anonymisation will usually not exist.

7.3.2.2 Sanctions

If a processor is commissioned who wants to process recognisable data for his own purposes in an incompatible manner, the client runs the risk of violating the requirement to **carefully select** the service provider (Art. 28 (1) GDPR). This violation can be sanctioned with a fine of up to 10 million euros or up to 2% of the worldwide annual turnover according to Art. 83 GDPR.

By processing data for **its own purposes** in violation of instructions, the processor itself becomes the controller (Art. 28(10) GDPR). If there is no legal basis for this data processing, it is **unlawful**. Unlawful data processing can be sanctioned with a fine of up to 20 million euros or up to 4% of the annual worldwide turnover according to Art.83 GDPR.

Under certain circumstances, e.g., in the event of failed or insufficient anonymisation, compensation for the **material or immaterial damage** suffered by the data subject may also be considered pursuant to Art. 82 GDPR, for which the controller and the service provider are **jointly and severally** liable.

7.4 Anonymisation within the group of companies

If anonymisation is carried out within a **group of companies** or if anonymised data is passed on, the general requirements for the involvement of third parties must be observed. In this respect, the GDPR does not recognise a genuine **group privilege**. When

examining whether anonymisation is sufficient, it will depend on which means are available to a parent company, for example, to identify a natural person or how likely it is that they will be used. In principle, such means can also be an instruction issued by a parent company to a subsidiary regarding the release of data.

8. Selected application classes

In chapter 6.3.1 selected procedures for anonymising personal data were presented. The aim of applying anonymisation procedures to personal data is to generate data that can be used for certain applications, but no longer have any personal reference. This is the case when the generated data can no longer be used to re-identify data subjects without disproportionate effort. The properties of anonymised data differ depending on the use case.

In this chapter, the selected application classes **'anonymisation as deletion'**, 'anonymisation **during disclosure'**, 'anonymisation **for training algorithms'** and **'anonymisation for testing software' for the** use of anonymised data are presented. The primary aim is to illustrate the possibility of providing anonymisation from personal data that is precisely tailored to the application in practice. Furthermore, possible limits of the usability of anonymised data are to be shown.

One of the aims of anonymising personal data is to minimise the risk of unauthorised access to personal data or the re-identification of data subjects by unauthorised persons, so-called attackers. Therefore, when designing the anonymisation solution, a possible approach of an attacker is taken into account (see also chapter 6.2). To illustrate this, the structure of the examples presented below includes a brief description of the possible **attacker** and the **personal data** under consideration to be protected. In the example, the desired **use or processing of** the personal data is first described. Then the **anonymisation solution** for the example is presented. The **properties of the anonymised data** are then described and evaluated in terms of processability for the planned application.

8.1 Anonymisation as deletion

The deletion of personal data after use for a specific purpose is one of the obligations of the data controller. If deletion is regarded as the removal of the reference to a person, anonymisation can be regarded as deletion. It must be noted here that the personal data must be removed from the data processing system immediately after anonymisation has been created and the legal basis for data processing has expired. The remaining anonymised data no longer have any personal reference. Accordingly, their use is not bound by data protection regulations.

One advantage over completely removing the data from the data-processing system is the possibility to use the anonymised data independently of the processing purpose of the original personal data. The examples described below are intended to demonstrate

this possibility. In addition, limits to the usability of data anonymised as deletion are to be shown by way of example.

8.1.1 Example: Keeping key data on applications

As is generally known, applications are to be deleted after the recruitment decision has been made and the limitation period of claims, especially from the AGG, has expired. The deletion period is thus a few months. HR managers are faced with the task of having to measure the placement success as well as the quality of the applicants for years. If the application data were completely deleted, such measurements could not be made or could only be made for very short periods of time.

Instead of completely deleting the data record, anonymisation is an option. In practice, anonymisation is equated with the deletion of obviously identifying characteristics such as name, address or contact details. A combination of "skills", e.g., in connection with stations in the curriculum vitae, can enable re-identification with the help of profiles in social networks. Therefore, it should be examined whether further anonymisation techniques need to be applied after deleting the obviously identifying characteristics.

An **attacker** may be interested in sensitive information of the applicant such as work history, details of desired salary and severe disability.

The **personal data considered** include date of birth, place of birth, school career, language skills, nationality, address and possible information on a severe disability.

If the personal data has been anonymised in such a way that it is to be considered deleted, the allocation of individual information to individual applicants is no longer possible without a disproportionate effort. As a result, certain key data about an applicant can no longer be collected from the anonymised data. However, there are possible benefits of anonymised data compared to deletion for information that can be determined from suitably anonymised data. **Examples** of this are listed in **Table 3** below.

Personal data	Anonymisation
Years of the applicants	Average value over all applicants in year x for several years. (A)
Origin	Generalisation of origin by region (e.g., Central Europe, Middle East, North America) (V)
Language skills	Average of the number of languages mastered across all applicants in year x for several years. (A) Average of the number of applicants who were proficient in English at level C1; in year x for several years. (A)
Educational level	Average values of school-leaving qualifications achieved in year x; Number of applicants with a Bachelor's degree. (A)
Final grades	Average grade, broken down by years and degrees. (A)
Residence	Region of residence, e.g., Rhineland instead of Bonn. (V)
Severe disability	Proportion of severely disabled applicants in year x. (A)

Table 3: Examples of personal data and the replacement of these with appropriate generalisation (V) or aggregates (A).

The anonymisation solution here consists of applying procedures for calculating averages or generalisation and then completely removing the originally recorded personal data. From the anonymised data, statements can be derived about the totality of applicants in a certain period, e.g., year x. Furthermore, the allocation of information to a specific applicant is no longer possible.

8.1.2 Example: Quality analysis of the customer service of an electrical retailer

An **electrical retailer** traditionally offers the sale, maintenance and repair of electronic equipment. The **customer service department of the electrical retailer** wants to know the needs of customers who contact him particularly frequently because of problems with the previously purchased devices. For this purpose, he would first like to save all the data collected about a customer's call and analyse it later: The transcribed customer conversation, the previously purchased appliance, the name, age and, derived from the

conversations, the customer's affinity for technology and level of education, frequency of use of the appliance etc.
In order to be able to analyse the customers' data in a legally secure way, the electrical retailer obtains their consent to analyse the conversations. He notices that 85% of the customers do not give this consent. Therefore, he is obliged to delete at least the personal data of these customers after the end of the customer service process associated with the conversation. The starting point of the customer service process is marked in a ticketing system by an open ticket. The closing of the ticket by a customer service employee means the end of the process.

Solution: After closing the ticket associated with the conversation, the customer service enriches aggregates of the previously collected personal data. The remaining data is deleted.

An **attacker** may be interested in details from the customer conversations from which a motivation to switch to another dealer or customer service can be derived. For example, he may also be interested in the number of orders and the number of enquiries to the customer service department.

The **personal data considered** include

- Time and duration of the interview
- The text of the transcribed conversation
- Name of the customer
- Age of the client
- Indicators of the customer's affinity for technology and level of education, recorded e.g., by personally stating the highest degree or by the correct use of technical terms.

Other information, such as the device purchased, may also be considered personal. This is subject to the assessment of a data protection officer responsible for the electrical retailer.

The **use of the data** includes the following at the time of opening the ticket:

- Identifying which customer service department should be involved in resolving the customer's query and forwarding information to them.
- Callback and appointment
- Receipt and repair of defective units; allocation to individual customers; return of the repaired unit to the customer.
- Refund or partial refund of purchase prices

After closing the ticket, the use of the data is restricted to those analyses or processing based on aggregated data without direct personal reference.

The data is already anonymised before the ticket is closed. This is done by enriching aggregate values by adding the information determined in the course of processing the customer transaction. If necessary, time windows can be defined within which accruing data is aggregated. In this way, more precise statements can be made about the needs

of customers and their behaviour over certain periods of time. It is also possible to create value pairs of aggregates. This enables, for example, two-dimensional categorisations of the information derived from personal data. After closing the ticket, the original personal data is removed from the system.
Examples of anonymised data replacing deleted personal data are listed in **Table 4** below:

Personal data	**Anonymisation**
Time and duration of the interview	Duration: Three categories: short duration (<5 minutes), medium duration (<10 minutes), long duration (>=10 minutes). Time: Time slots: morning, noon, afternoon, evening. If required, two-dimensional categorisation: duration x time.
The text of the transcribed conversation	Type: General complaint, late delivery, defective goods, warranty case, guarantee case, repair request, ...If required, multi-dimensional categorisation (e.g., late delivery, defective goods).
Name of the customer	removed
Age of the customer, e.g., by recording the date of birth	Age category: 18-25, 26-35, 36-45, 46-55, 56-65,>65.
Purchased device	Remains
Indicators of the customer's affinity for technology and level of education	Average of clients using technical terms correctly. Average of clients who have achieved at least the Abitur.

Table 4: Information that should be available after deletion without preserving the person reference.

8.1.3 Example: Website statistics

The operator of an online shopping platform offers the possibility to order and deliver goods. To simplify orders, customers can create profiles through which orders can be processed in a simplified way without customers having to re-enter their data each time they place an order. Furthermore, it is possible to mark products found in order to find them again more easily later, as well as to set preferences for new offers from certain categories.

Attackers in the area of website statistics are located in the internal area in this example. Of course, the leakage of customer profiles to external companies or institutions would also be possible, but such examples are dealt with in the chapter on 8.2 chapter.

In this scenario, the main danger is an unauthorised use of the data for other purposes. For example, the data is no longer used purely to process orders, but is used in further analyses, which should ultimately lead to an increase in sales. In addition, users can also be identified and orders assigned retrospectively.

The personal data considered is primarily shipping information such as the full name, the associated delivery and billing address and the telephone number for contacting in the event of a problem or in the case of larger shipping deliveries. In addition, the e-mail address for account management and payment information are required. This includes the type of payment in addition to possible numbers such as the credit card number.

Active use of the platform results in information on products or product groups that are of interest to the respective person.

The **data** shall initially be **processed** within the scope of the intended purpose approved by the customer, i.e., order processing and services to simplify the use of the platform. If the customer objects immediately or later to further use of the data, this withdrawal of rights must be taken into account.

The question remains open as to whether the deletion of data is absolutely necessary and which data this should concern. The problem here is that all usability of the data would be lost.

In this case, **anonymisation solutions** can largely maintain the usability of the data. If categories are defined in advance, it is possible to use them to categorise the users. Thus, the name can be removed and only the gender can be stored. If age is specified, it can be generalised or categorised as "youth" or "young adult", ..., "senior". Addresses can be generalised to (larger) postcode areas. Telephone numbers could be removed or generalised analogous to postcodes and payment information could again be categorised. These would be, for example, "direct debit", "surname" or "instant bank transfer".

The **properties of the anonymised data** continue to fulfil the main aspects of usability, but if quality criteria such as *k-anonymity* or others are observed, the personal reference is to be considered removed and the data is no longer subject to the GDPR.

However, desired statistical analyses remain possible on the anonymised data. For example, it can be determined whether people from certain geographic regions or age groups prefer certain product categories or tend to choose higher-value products. This can be used to optimise advertising or product displays.

Table 5 below shows a simplified excerpt from the orders of a fictitious bicycle dealer, which concentrates on the essential characteristics for understanding the procedure.

The name, gender, address, age, email, payment method with further information on possible return transfers and the purchase are recorded.

As part of a **marketing campaign** and to **optimise the online shop,** the mail order company would like to know which people should be shown which products. For this purpose, he creates table 5, which still enables these analyses, but no longer has a reference to persons and also fulfils 2-anonymity, i.e., k-anonymity with k=2.

This allows us to determine that in the area of the generalised postcode 5311*, low-priced bicycles tend to be sold, while in the area of 5322*, high-priced bicycles are purchased. These customers were equally male, while female or diverse tended to prefer slightly cheaper products. Based on the categorisation by age, it can be seen that young adults prefer high-priced sporty bikes, while adults tend to prefer normal to low-priced ones. Seniors, on the other hand, buy bicycles that are characterised by their comfort.

With more characteristics in the table and a larger data basis through more table entries, these analyses can still be significantly expanded and refined. However, this example already shows that market analyses remain possible even on anonymised data.

Name	**Gender**	**Address**	**Age**	**E-mail**	**Payment method**	**Purch-asing**
Max Sample	d	Sample path 1 53115 Bonn	40	mus-ter@bsp-mail.de	Giropay DE77 8765 7896 8907 8970 00	"Wire bike" 799€
Luise Müller	w	Forest path 7 53115 Bonn	65	lm379@bspmail.de	PayPal DE86 9872 1234 5674 5678 32	"Holland Wheel Su-per" 577€
Maximilian Müller	m	Main street 3 53229 Bonn	21	max@bspmail.de	Instant bank transfer DE65 2853 4637 7531 8953 55	"Sport bike fix" 1699€
Edgar Michels	m	Larch path 3 53225 Bonn	19	em@bsp-mail.de	PayPal DE53 5674 1842 5953 0000 00	"Mountain Bike Ro-deo" 2099€

Paris Gebhart	d	Meadow path 3 53117 Bonn	38	pa-ris777@bspmail.de	PayPal DE22 0056 7431 8642 4533 33	"Budget bike" 299€
Marianne Henschel	w	Sample path 12 53115 Bonn	66	hen-schel@bspmail.de	Cash on delivery	"Wheel Comfor-table" 649€

Table 5: Database of a bicycle dealer

Name	**Gender**	**Address**	**Age**	**E-mail**	**Payment method**	**Purch-asing**
-	d	5311*	Adult	-	Giropay	"Bike 7" 799€
-	w	5311*	Senior	-	PayPal	"Holland Wheel Su-per" 577€
-	m	5322*	Young adult	-	Instant bank transfer	"Sport bike fix" 1699€
-	m	5322*	Young adult	-	PayPal	"Mountain Bike Ro-deo" 2099€
-	d	5311*	Adult	-	PayPal	"Budget bike" 299€
-	w	5311*	Senior	-	Cash on delivery	"Wheel Comfor-table" 649€

Table 6: Anonymised database of a bicycle dealer

8.2 Disclosure of anonymised data

When data are transferred to third parties, depending on the purpose of the processing, there must no longer be any personal data that allow the identification of specific individuals. The examples presented here present situations in which data is passed on and make suggestions about the techniques that can be used in the process, through which a desired processing or use is still possible without having a clear personal reference.

8.2.1 Disclosure of payrolls

The salary structure of different companies from the same industry helps HR managers to classify employees' salary wishes and to assess the attractiveness of the salary structure in the market. Providers of salary comparisons collect salary data directly from interested persons or request complete salary lists with characteristics such as qualification, position, industry, place of work, age, salary components, gender directly from companies.

The legitimising legal basis for the disclosure of personal salary lists is regularly lacking. It makes no difference to the lack of a legal basis whether the disclosure is made against payment or free of charge. A balancing of interests does not apply, as the principle of confidential treatment of personnel data outweighs the employer's legitimate interests. It is conceivable that employees may consent to the disclosure. As not all employees agree as a rule, only a few data records can be transferred.

Anonymisation before disclosure could be legitimised by a balancing of interests, as anonymisation significantly reduces the depth of intrusion into the personal rights of employees. In order to be able to use the balancing of interests, the purpose of anonymous disclosure must already have been determined when the data concerned were collected and must have been made known to the employees as part of the data protection information. Otherwise, it must be checked whether the new purpose of the disclosure is compatible with the original purpose. The verification procedure is described in Article 6 (4) of the GDPR. If the check is positive, all employees concerned must be informed of the new purpose.

In order to compare salary structures, the relevant influencing factors must be recorded. These include age, position, qualification and gender of employees. Even the position is person-related, especially in the case of managers. As a rule, there is only one head of HR. However, the combination of age and gender can also be person-related, for example, if there is only one woman among the 30-40 year olds.

The strategy of deleting all features that can contribute to an identification would in fact delete (almost) all features. No usable data set would remain. In order to achieve anonymisation, it therefore makes sense to switch to other anonymisation methods. An example of this would be the generation of synthetic data on the basis of the statistical properties of the personal data underlying the synthesis model.

8.2.2 Example: Sharing sales figures by product category

A mail order company offers goods from different product categories. She finds that customers from different buyer groups are willing to pay different prices for the same products. She now wants to capture this information and market it to third parties to strengthen her own sales.

First, she starts by determining the tolerance values of the respective user groups. To do this, it randomly varies the prices of its sold products and records some additional parameters of these and the associated buyers when the purchase is concluded.

These parameters include the following personal and associated data: For a billing address and a purchased product, the price, the time of purchase with date and time, the devices used, browser and internet provider with IP addresses are stored.

Attackers in this scenario can be seen as both internal and external parties who may have an interest in detailed profiling. Use could then be used for unwanted personalised advertising, phishing or subsequent identity theft. Even if no names of buyers are stored during the collection process, it is possible, especially for an internal attacker, to use existing data ("background knowledge") to clearly assign data sets to a person again. The buyer of such a data set could also aim to re-identify a person or the entire data set. He could then not only adjust his own prices - the original goal - but also compare the received data with his own purchase history in order to re-identify persons based on combinations of recorded additional characteristics.

Since the collection of this data for a purpose other than the original purpose of order processing is already not permissible, anonymisation solutions must be applied here that still allow the original use as a pricing mechanism. Therefore, the mail order company does not store the data directly when a purchase is concluded but reduces it to categories while removing all direct personal characteristics. In doing so, it takes particular care to ensure that there is no match for any of the characteristics with less than *k* other records. A schema for the data collected above could now look like the one shown in the table below, with the last line repeated accordingly often for all variants of the upper parameters.

Parameter	**Transformation**	**Example value**
Product	Replace with product category with similar purchasing behaviour	Samsung Galaxy Tab S7 Upper class tablet, >2020
Invoice address	Replace with location range covering *k* other entries of the table, e.g., with Open Location Code or a more general city (part of) specification.	Trankgasse 2, 50667 Cologne 9F28WXR4+ or: City centre, Cologne

Date, time	Replace with relevant categories	Evening, autumn, no holiday, no holiday day
Device	Replace with category	iPhone 12 Pro Apple smartphone
Browser	Deletion of specific features	Google Chrome 104 (Windows) Google Chrome, latest version
IP address	Anonymisation by replacing with the corresponding ISP	84.128.17.3 German Telekom
Price	Indication of the possible price difference in %.	700 € +5 % of the OVP

Table 7: Anonymisation of personal data in preparation for disclosure.

8.2.3 Example: Anonymous matching of leaked access data

An online service where registered users log in with an email address and password wants to improve the security of its platform. To do this, it wants to compare the credentials used to log in its users with a list of stolen and published credentials (called leaks) in order to warn its users and prevent an unauthorised attacker from using these credentials to log in to its service.

Lists of such stolen credentials published or sold on the internet are collected by different service providers for matching. A very simple way of matching would be for the online service to send its service provider the access data used by its users - the email address and the unencrypted password - when they register, so that the service provider can compare them with its list. In doing so, the service provider would now have direct access to personal data that the original user has not consented to the processing of, and which could additionally be linked to each other across services and profiles could be created about a customer.

In the EIDI research project led by the University of Bonn, a protocol was developed and implemented in which no personal data is transmitted to the service provider in plain text, but a reliable statement can still be made as to whether the access data is in a leak database.

During a login attempt, the online service to which the user wants to log in uses a cryptographic hash procedure to convert the user's email address into a random-looking string of characters. Due to the characteristics of such functions, it is still possible to assign the e-mail address to a complete hash, which would endanger anonymity. If a hash generated in this way is now truncated and only a few digits (called a prefix) of it are passed on to the leak service provider, k-anonymity can be implemented here. The length of this prefix is chosen in such a way that there are always at least k other entries

in the database that have this prefix, and therefore k-anonymity is granted. The leak service provider can now search its database for entries with hashed e-mail addresses that begin with the prefix received.

Now, to compare the passwords stored in the leak service provider's database, both sides use an encryption method based on elliptic curves to encrypt the chosen password with two keys. First, the online service uses its secret key to encrypt the password and sends this ciphertext to the leak service provider, who encrypts it a second time with its secret key and sends it back. Due to the mathematical properties of the chosen encryption system, the online service can now remove "its" encryption and receives the password as if it had only been encrypted by the leak service provider, but without the latter being able to read it for encryption.

The leak service provider can now send the hits of the query in the list of leaks encrypted with its secret key to the online service, which can then compare the k-anonymised hits with the user's encrypted password without knowing it. This comparison is now possible because the same key was used in each case. In the event of a hit, the online service can then take further measures to protect the account of the user concerned and inform him or her.

8.2.4 Example: Fuel consumption of vehicles

Modern cars of newer model years are characterised by an increasing increase in comfort and safety. For example, multi-function steering wheels are installed that make the car more pleasant to operate or automate processes such as switching on the headlights and windscreen wipers. At the same time, more safety systems are being installed, such as a larger number of airbags.

In addition, there are safety systems such as the eCall[22] , which automatically triggers an emergency call in the event of an emergency and at the same time can transmit the position of the vehicle via mobile phone and GPS tracking. In addition, the EU requires so-called On-Board Fuel Consumption Monitoring (OBFCM) for newer cars in order to monitor fuel consumption and compare it with the manufacturer's specifications.[23]

Attackers of this data transfer can be of many kinds. For example, insurance companies could be interested in identifying drivers, as this can generate interesting information about the driver's driving behaviour.

[22] https://www.verbraucherzentrale.de/wissen/reise-mobilitaet/unterwegs-sein/ecall-so-funktioniert-das-automatic-emergency-system-in-car-32100 (last accessed 28.11.2022).

[23] https://www.adac.de/rund-ums-fahrzeug/ausstattung-technik-zubehoer/assistenzsysteme/obfcm/ (last accessed 28.11.2022).

Personal data considered include the usage profile and driving style, as analyses by the ADAC show[24][25] .
Broken down individually, the extent of the data storage becomes clearer. In addition to storing the kilometres driven separately for city, country and motorway and individual routes, the fuel consumption and speed are also recorded to the second. In addition, GPS data and other important vehicle data are stored.
The driving style is analysed based on the number of belt tightenings, engine speed and temperature or traction battery data and use of the different operating modes, e.g. sport mode.

Processing of the data can be done for the intended purpose required by the EU regulation. Thus, an analysis of fuel or electricity consumption by city, country and motorway makes sense, as these must also be provided in the consumption data required by manufacturers.

The further recording of the driving style allows conclusions to be drawn about the use of the car. Thus, a sportier driving style is expected to have higher consumption than defensive driving. This could be used to correct or classify the consumption values.

The data also serve the convenience of the driver, as the consumption values are transmitted to the display and thus do not necessarily have to be determined by documentation and calculation at fuel stops, as was the case in the past.
The localisation function is necessary for the eCall and can even save lives in this case.

Anonymisation solutions are obviously unavoidable when looking at the recorded data, as delivering the data into the wrong hands would create a transparent driver whose place of residence, place of work and other locations could easily be determined from the movement profile.
The mileage can be categorised by "city", "country", "motorway", while the average consumption is assigned to these categories. The vehicle type must be supplied for comparison. The records of individual driving distances are suppressed, as is GPS data.

For further analysis, values of the recorded parameters can be set, according to which drivers can be categorised as "speeders", "sporty drivers", "average drivers", "defensive drivers".

The **properties of the anonymised data** can still serve the required statistical analyses and provide the required comparison with the consumption data of the manufacturers. Even a separation into city, country and motorway remains possible, as well as a further analysis of consumption that takes into account the driver's driving profile.

[24] https://www.adac.de/rund-ums-fahrzeug/ausstattung-technik-zubehoer/assistenzsysteme/obfcm/ (last accessed 28.11.2022).
[25] https://www.adac.de/rund-ums-fahrzeug/ausstattung-technik-zubehoer/assistenzsysteme/daten-modernes-auto/ (last accessed 28.11.2022).

8.3 Anonymisation when training algorithms

When anonymising training data, the correlations in the data that represent the patterns to be recognised must not be changed. A prerequisite for the selection of suitable methods for anonymisation is therefore that it is known which correlations in the data are essential for the patterns to be recognised.

The training of algorithms is problematic from several perspectives in terms of data protection law. This was the conclusion of a study[26] by the German Association for Information Technology, Telecommunications and New Media (Bitkom). In principle, the precise training of a model requires a broad database to provide sufficient information. Otherwise, there is a danger of creating a model that is too coarse. If the algorithm is supposed to recognise cars in later use with the help of the learned model and only images of SUVs were used for training, it could happen that buses are also classified as cars, while small or sports cars are left out.

In the context of data protection or the protection of company secrets, however, a consolidation of large amounts of data is often problematic, as these become known in the company offering the creation of artificial intelligence models and thus end up in foreign hands. This is where federated learning comes in, for example.

8.3.1 Example: Federated Learning

In principle, **federated learning** means that the data can be used for training as well as remain with the respective data owner. The service provider first creates an initial model, which it passes on to its partners. Using their respective data, each partner now tests the received model and informs the service provider how the parameters should be changed to improve the model. From all the feedback received, the service provider now calculates an overall update and applies it to the previous model, which is now distributed again. The analogy to the already introduced round-based optimisation of machine learning is immediately recognisable here. The advantage here, however, is that the data remain with the respective owner. [27]

Variants of federated learning are currently being tested in research. Unfortunately, no practical applications are known so far.

However, the application of federated learning is not yet sufficient from a data protection perspective. Based on its answers to posed questions, an algorithm trained on real data can reveal information about the training data. Thus, answers to data sets used in training may differ (minimally) from answers to novel data sets, leading to the discovery of personal data.

[26] Anonymisation and pseudonymisation of data for machine learning projects - A handout for companies, Bitkom 2020.

[27] Anonymisation and pseudonymisation of data for machine learning projects - A handout for companies, Bitkom 2020.

8.3.2 Example: Differential Privacy

In order to solve the problem of 8.3.1 **Differential Privacy** can be applied to solve the problem of revealing personal data by observing changes in the evaluation of data sets. On the one hand, differential privacy can be applied to original datasets before they are used for training. It is important to note that the resulting datasets further ensure good model quality. A "trade-off situation" arises. On the other hand, the application of differential privacy mechanisms to the outputs of the algorithm is possible.

These are also current research topics whose suitability for practice has yet to be demonstrated.

8.3.3 Synthetic data

Another approach is to dispense with real data altogether when training algorithms. **Synthesised data** can be used for this purpose. A good synthesis model is of elementary importance here, which may itself have to be trained. This involves algorithmically generated synthetic data. If, for example, it is a question of predicting diseases for certain population groups, care must be taken to ensure that statistical conditions remain intact. The danger is that clusters of certain diseases that do not occur in reality render the model useless. Moreover, even the use of synthetic data does not fully protect against the discovery of individuals who were once used to create the synthetic model. Here, too, the practical suitability in terms of data protection and, above all, model accuracy must first be demonstrated.

8.4 Software testing

The development of software involves extensive testing. One of the purposes of conducting the tests and evaluating the test results is to ensure that the desired functionality is correctly implemented in the software. So-called test data are used to test software. This is data that is at least similar to the data to be used in the productive environment. If, for example, new software is being developed that is already in use, real data can be used to test the newly developed software. However, this is subject to severe limitations under the GDPR if the real data is personal. In such cases, it is advisable to use synthetic data. In order for the test result to represent the expected behaviour of the software at the time of use sufficiently accurately, the test data must be sufficiently similar to the real data with regard to certain properties. For this purpose, the development team carries out an assessment of the most important properties of the real data with the expected effect on the behaviour of the software. In order to be able to generate the synthetic data as accurately as possible, the properties assessed as important are reproduced in the modelling of the synthetic data.

When generating test data taking into account the characteristics of real personal data, it is important to ensure that the test data does not replicate the real data in such a way that it is possible to re-identify data subjects by using the test data. In the following, the generation of test data is illustrated by examples.

8.4.1 Software update/migration

The personnel management system PVS, which has been used in company X for many years, has all the functionalities X wanted. These include e.g.

- the bookkeeping with payroll
- the payment of salaries
- the management of the employees' master data
- the holiday planning.

In addition, X's staff are well acquainted with the use of PVS. PVS is distributed and maintained by company Y.

It is now known that PVS has serious security gaps. Company Y reports that these gaps cannot be remedied without a fundamental revision of the system's code base. X would like to continue using PVS and therefore decides to commission Y with the revision.

The PVS is being completely redeveloped by Y. The functionalities originally available in the PVS are replicated. The user interface is also modelled on that of the old PVS. Only modern, comparatively secure technology is used.

Now Y wants to test the newly developed PVS in different phases. For this purpose, all processes are to be simulated which, according to the documentation, were carried out on the old system in the last two years using the personal data of the customers and employees. Some examples are listed below:

- An employee A was hired as of 01.10.2021. His income tax class is 1, he lives at Bonnweg 123 in Cologne and is a Russian citizen. His salary is EUR 50 000 p.a.. The position is limited until 01.12.2024.
- A one-off Corona payment of EUR 1000 gross was transferred to all employees on 01.12.2021.
- Employee A has registered his annual leave 2022 for the period 25.03.-06.05.
- The average salary of one of the 50 employees was EUR 48,000 p.a. in 2021.
- The managing director had an annual salary of EUR 120,000. He thus achieved the highest annual salary.
- The average turnover generated by one of the 250 customers in 2021 was EUR 20,000.

For reasons of data protection, company X cannot pass on the personal data of the customers and employees to company Y.
By accessing the data, a potential **attacker** could learn details of the private life of the person concerned, which could constitute a serious interference with his or her right to informational self-determination.

This attacker could, for example, act in the role of the software tester and gain unauthorised insight during the validation of the test data. Since companies X and Y are both located in the Cologne/Bonn area, it is also not unlikely that the employees and customers know each other personally.

The personal data of an employee considered in the PVS include name, address, income tax class, nationality, annual salary, date of employment contract and period of annual leave.

Other data include the average profit per client, the average annual salary and the timing and amount of the Corona one-time payment.
The **planned processing** of the data includes actions to be replicated as part of the testing of the new PVS. These are

- Creating a record for a new employee: This is where the personal data is collected and stored.
- Entering the holiday planning of an employee: Here, the period, the number of days of holiday is saved as an additional entry assigned to the employee.
- The monthly toasting of the transfer of salaries
- Triggering the transfer of the Corona one-off payment
- The calculation of the average annual salaries
- The difference between the highest annual salary and the average annual salary.
- The calculation of the average of the sales of a year per customer from the sum of the sales generated in a year.

In the production-ready PVS, the personal data would be extracted from the corresponding database (here PVS-DB) in order to then be used in individual actions or calculations. In order to simulate this situation, Company Y needs data that is similar to the original personal data, but not identical to it.

An **anonymisation solution in** this case is the artificial generation of test data that replicates attributes of the personal data. For this purpose, a database is created with the same attributes of the PVS-DB. The values are set using artificially generated data collections. Suitable examples for the generation of artificial data instead of personal data are described below.

- Name: Publicly available lists of common family names can be used here.
- First name: Lists of names from registry offices can be used here. These are lists containing suggestions for female and male first names.
- Address: The following pattern can be used here, taking into account the preservation of the format of the PVS-DB entries: first name, last name, street, current line number, random five-digit number as postcode, random selection of a city from the Cologne/Bonn area.
- Income tax class: Random selection of tax classes 1-6
- Annual salary: Random selection of a value higher than the current minimum wage.
- Entry date: Random date
- Planned exit: Random date.

Properties of the anonymised data: The result is a database whose entries are anonymously created data based on personal data, which is syntactically and structurally similar to the personal data but has no reference to an identifiable natural person. This prevents combinations of name, date of birth, nationality and address that occur in the real data from being transferred "unaltered" into the test data. This makes re-identification of data subjects from the test data impossible.

> **Note:** The calculations on the anonymous data are analogous to the calculations performed on the original data.

8.4.2 Functionality tests

Looking now at the previous example, testing the functionality of privileged users of the system with certain permissions may also be necessary. The generation of the test data described in chapter 8.4.1 is not sufficient for this functionality, since the level of access authorisations is to be examined. However, the productive data of the user accounts including login credentials and entries of the authorisations are also **personal**. This means that use for purposes other than the productive system is regularly not permissible. In particular, passing on this information to Company Y is not permissible. An **attacker** could use the user accounts for unauthorised access to the PVS and thus gain access to a large amount of sensitive information.

An **anonymisation solution** that makes it possible to test this part of the software's functionality involves providing a test system that does not itself have access to real personal data of the PVS. On this test system, user accounts can now be created with the various authorisations that the real privileged users have. However, these data are enriched with synthetic data at the positions where personal, possibly sensitive information attributable to a natural person would be found in the real data. This user data can now be used in tests to check the correct functionality of the user authorisations and their settings in the PVS.

9. Other legal requirements

9.1 Documentation requirements

The measures taken and the relevant influencing factors for determining an appropriate anonymisation procedure must be **documented**. This can be done either by means of an independent **anonymisation concept** or by means of a general description within the framework of the **presentation of technical-organisational measures for** a processing activity (see chapter 9.2). The procedure should be **transparent** and **comprehensible** to outsiders. The implementation of this requirement in the documentation can be checked using the following questions:

- Can the effectiveness of the procedure be verified?

- Can the measures applied in the procedure be verified with regard to their implementation?
- Can compliance with the measures applied be evaluated?

All persons involved in the implementation should be able to understand the process or measure and implement it according to the defined specifications.

It is advisable to base the documentation on the **procedure model for anonymisation** (see chapter 9.5).

9.2 Record of processing activities

Every controller within the meaning of Article 4(7) of the GDPR must keep a **record of its processing activities (RoP).** A processing activity can be understood as a sequence of different processing steps that serve at least one overarching purpose (e.g., applicant management, personnel management or accounting). The processing steps are related to the higher-level purpose in terms of content. The technical aids used, e.g., software programmes, are not to be taken into account in the delimitation of processing activities.

The anonymisation of personal data is basically to be classified as **processing** (cf. chapter 4.2). However, anonymisation regularly does not pursue its **own purpose**, but is a processing step that serves a **higher-level processing activity** (e.g. statistical evaluation of user behaviour). It is therefore recommended to describe anonymisation within a processing activity either as a technical-organisational measure or to refer to a separate anonymisation concept. Anonymisation of personal data is therefore not a separate processing activity with regard to the record of processing activities.

With regard to the legal requirement to carry out a **data protection impact assessment** for processing operations likely to result in a **high risk to** the rights and freedoms of natural persons, a risk analysis for the rights and freedoms of data subjects should be carried out for the respective processing activity. Anonymisation does not per se result in a high risk to the rights and freedoms of natural persons. Other facts must be added which, taken as a whole, lead to a high risk.

9.3 Data protection information

Pursuant to Art. 13 of the GDPR, the controller must **provide** the data subject with comprehensive information on the purposes of the data processing, the legal basis and recipients or categories of recipients, among other things, already at the time of data collection. Pursuant to Article 14 of the GDPR, there is also an obligation to provide this information if the data controller has not collected the data directly from the data subject, for example from third parties or from public sources such as the Internet.

Since the process of anonymisation constitutes data processing (see chapter 4.2), information about planned anonymisation must therefore be provided at the time of data

collection or notification. According to the wording of Articles 13 and 14 of the GDPR, this information must also include the purpose of the processing, i.e., the anonymisation. As the examples below show, these own purposes can be of a statistical nature, e.g., for activity, purchase or payment analysis. The legal basis for the permissibility of anonymisation for these purposes is usually the permissibility of the **balancing of interests** pursuant to Article 6 (1) (f) of the GDPR.

The processing operation of forwarding already anonymised data to third parties is no longer subject to the GDPR. Information about specific **recipients or categories of recipients is** therefore no longer required. Data sharing on the basis of the EU law is thus also possible.

A data protection information pursuant to Art. 13 GDPR on the processing purpose of anonymisation could read as follows: "In addition to business purposes, we anonymise your data for statistical and data sharing purposes based on the EU legal acts".

If data has been collected without prior data protection information on the intended anonymisation pursuant to Art. 13 of the GDPR, e.g., in the case of **older data files**, Art. 6 (4) (e) of the GDPR permits **further processing** if suitable safeguards are in place. According to the wording of the GDPR, this can also include encryption or pseudonymisation. This standard was created in order to enable big **data analyses** under the GDPR that are open-ended with the aim of pattern recognition or the generation of new (personal) information and to that extent without a specific or with a changing purpose. Insofar as pseudonymisation can already suffice as a suitable guarantee for a use that changes purpose, anonymisation is a far more effective guarantee. Article 6(4) would thus also allow older data to be used after anonymisation for analysis purposes or for data sharing.

Whether information about the subsequent anonymisation must still be provided is assessed in accordance with Article 14 (5) (b) of the GDPR. This regulation is based on a **disproportionate effort.** Statistical purposes are cited as an example of this. With regard to the right of personality and data protection, the data subject is no longer at risk after anonymisation. Subsequent information about anonymisation does not create any added value for the data subject in terms of data protection law, but only generates a disproportionate effort.

9.4 Review obligations

Anyone who has anonymised data or uses anonymised data is obliged to **continuously check that** anonymisation is maintained.[28] To this end, they must check whether the personal reference can be restored. The implementation and the result of the check should be documented.

Audited:

- Is singling out possible?

[28] Opinion 5/2014 of the Article 29 Working Party on anonymisation techniques, WP 216, p. 4.

- Is it statistically possible to link the record of the same person with records from legally obtainable records on that person?
- Can values of a characteristic be derived with a significant probability from values of other characteristics in the dataset?

When checking whether data is anonymous, it is not a matter of certain or correct findings. Rather, it is sufficient that a reference to a person can be re-established with a **certain probability** or for a part of the data records.
If a controller comes to the conclusion that a person is identified or identifiable, the requirements of the GDPR for data processing come into play, including the question of the lawfulness of the processing. If there is no legal basis on the part of the controller, the personal data must be deleted. In addition, it should be checked whether the removal of anonymisation results in an **obligation to notify** the competent supervisory authority or whether data subjects must be informed (cf. Art. 33 and 34 of the GDPR).

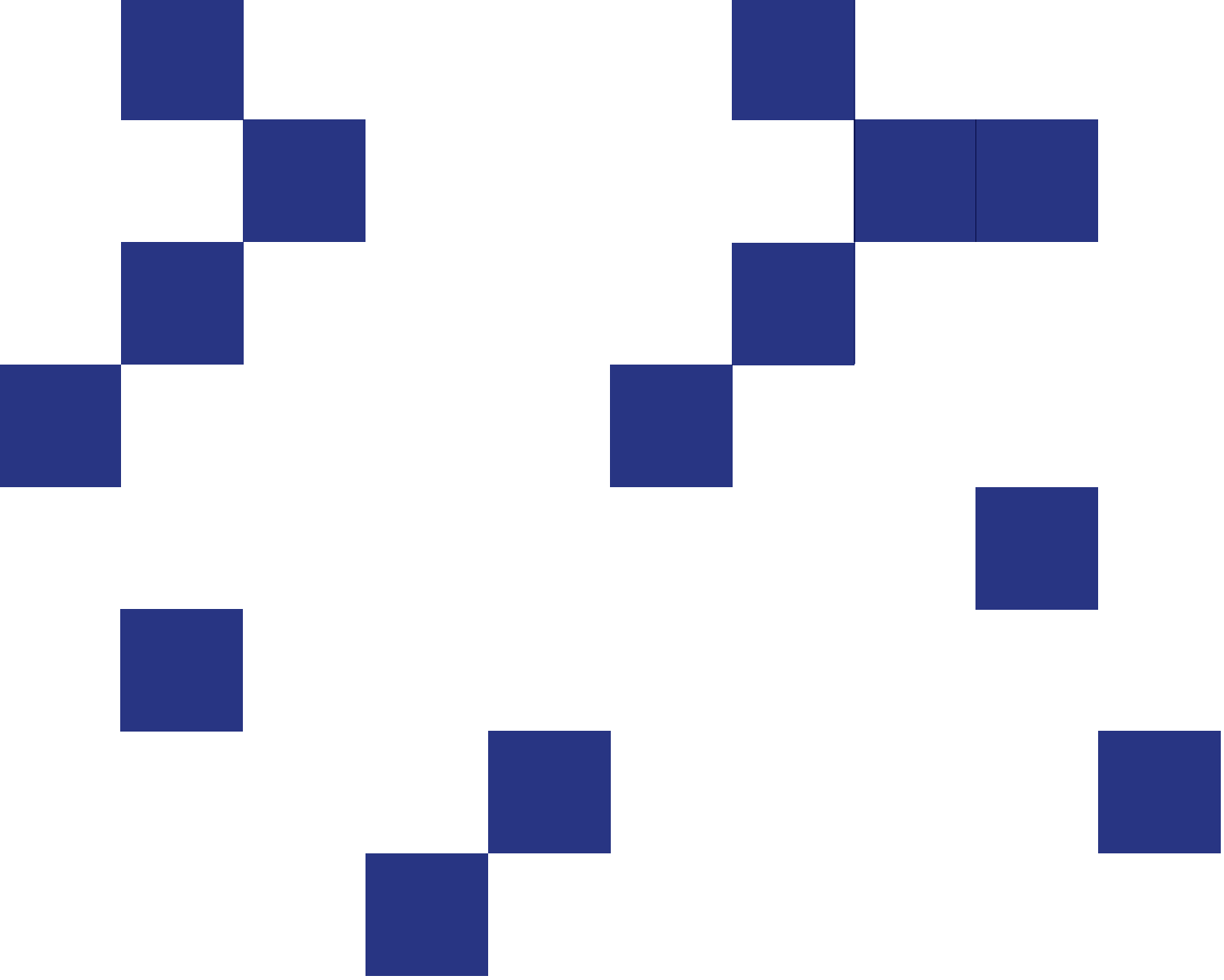
Note: Checking the storage period of anonymised data can also be useful in individual cases. After all, a deleted anonymised data set can no longer be the basis for a possible re-identification of data subjects.

9.5 Procedure model for anonymisation

The following procedure is suitable for anonymising personal data:

Seq. Number	Measure	Chapter in the Guide
1	**Identify the legal basis for anonymisation (e.g. in the GDPR, in the BDSG or in a sector-specific law).**	4.2
2	**Ensure that the information obligations pursuant to Art. 13 and 14 of the GDPR are implemented.**	9.3
3	**Selection and determination of the appropriate anonymisation procedure**	
3.1	Type and risk class of personal data to be anonymised	6.1 and 8
3.2	Intended purposes of processing	6.1 and 8
3.3	Context of anonymisation	6.1 and 8
3.4	Expected number of records	6.1 and 8
3.5	Identify the statistical properties in the datasets that are needed, and which characteristics are relevant for these properties	6.3 and 8
3.6	Determination of the appropriate anonymisation procedure and its timing	9.1, 9.2 and 9.4

4	**Conducting the anonymisation**	
4.1	Removal of all direct identifiers (e.g., name, address, contact details, credit card number).	3.2 and 6.1
4.2	Removal of all unnecessary indirect identifiers (e.g., gender, physical appearance characteristics, age, postcode).	3.2 and 6.1
4.3	Carrying out one or more procedures of the • Randomisation • Generalisation or • Carrying out a procedure with synthetic data.	6.36.3.1
5	**Analysis of whether and which risks exist for restoring the personal connection**	6.1 and 6.2
6	**If risks exist, application of further procedures for anonymisation**	6.3.1
7	**Go through steps 4.1 to 4.3 until no more risks are apparent.**	6.3.1
8	**Check whether the required statistical properties have been retained**	6.3
9	**Document test and result**	9.1
10	**Use or share anonymised data**	
11	**Regularly repeat the test according to step 5, apply steps 6 and 7 if necessary and document according to step 9.**	



STIFTUNG
DATENSCHUTZ

Stiftung Datenschutz
foundation with legal capacity under civil law
Karl-Rothe-Straße 10–14
04105 Leipzig
Deutschland

T +49 341 / 5861 555-0
mail@stiftungdatenschutz.org
www.stiftungdatenschutz.org

Funded by the Federal Republic of Germany
represented by Frederick Richter (Chairman)

The work of the Stiftung Datenschutz is funded from the federal budget (BMJ section).

Federal Ministry
of Justice