Bundesministerium des Innern, für Bau und Heimat

Digital Gipfel Nürnberg 2018

Rolf Schwartmann / Steffen Weiß (Ed.)

Requirements for the use of pseudonymisation solutions in compliance with data protection regulations

A working paper of the Data Protection Focus Group of the Platform Security, Protection and Trust for Society and Business at the Digital Summit 2018

Rolf Schwartmann / Steffen Weiß (Ed.)

Digital Gipfel
Nürnberg

Requirements for the use of pseudonymisation solutions in compliance with data protection regulations

A working paper of the Data Protection Focus Group of the Platform Security, Protection and Trust for Society and Business at the Digital Summit 2018

Chairman of the Focus Group:

Prof. Dr. Rolf Schwartmann
Cologne Research Centre for Media Law - Cologne University of Technology
Member of the Data Ethics Commission of the Federal Government

Coordination:

Steffen Weiß, LL.M.
German Society for Data Protection and Data Security (GDD)

Members:

Prof. Dr. Christoph Bauer
ePrivacy GmbH

Patrick von Braunmühl
Bundesdruckerei GmbH

Dr. Guido Brinkel
Microsoft Germany GmbH

Susanne Dehmel
Federal Association for Information Technology, Telecommunications and New Media e.V. (BITKOM)

Philip Ehmann
eco - Association of the Internet Industry

Walter Ernestus
The Federal Commissioner for Data Protection and Freedom of Information (BfDI)

Nicolas Goß
eco - Association of the Internet Industry

Michael Herfert
Fraunhofer-Society for the Promotion of Applied Research

Maximilian Hermann
Cologne Research Centre for Media Law - Cologne University of Technology

Dr. Detlef Houdeau
Infineon Technologies AG

Angelika Hüsch-Schneider
Deutsche Telekom AG

Clemens John
United Internet AG

Annette Karstedt-Meierrieks
Association of German Chambers of Industry and Commerce (DIHK)

Daniel Krupka
German Informatics Society (GI)

Johannes Landvogt
The Federal Commissioner for Data Protection and Freedom of Information (BfDI)

Prof. Dr. Michael Meier
University of Bonn/German Informatics Society (GI)

Dr. Frank Niedermeyer
Federal Office for Information Security

Robin L. Mühlenbeck
Cologne Research Centre for Media Law - Cologne University of Technology

Jonas Postneek
Federal Office for Information Security

Frederick Richter, LL.M.
Data Protection Foundation (Stiftung Datenschutz)

Dr. Sachiko Scheuing
Acxiom Germany GmbH

Irene Schlünder
Technology and Method Platform for Networked Medical Research (TMF)

Sebastian Schulz
Federal Association for E-Commerce and Distance Selling Germany (bevh)

Dr. Claus D. Ulmer
Deutsche Telekom AG

Dr. Winfried Veil
Federal Ministry of the Interior

Dr. Martina Vomhof
German Insurance Association (GDV)

Benjamin Walczack
Independent Centre for Data Protection Schleswig-Holstein

Author:
Data Protection Focus Group of the Digital Summit

Chairman of the Focus Group:
Prof. Dr. Rolf Schwartmann

Kölner Forschungsstelle für Medienrecht

Technology Arts Sciences
TH Köln

Contact:
Steffen Weiß

German Society for Data Protection and Data Security

Heinrich-Böll-Ring 10
53119 Bonn, Germany

Phone: +49 228 96 96 75 00
Email: info@gdd.de
www.gdd.de

GDD
German Society for Data Protection and Data Security e.V.

Version 1.0, 2018

Foreword

Pseudonymisation as a bridge between informational and entrepreneurial self-determination

Machines should make life safer, lighter, more pleasant and longer. The human being, respectively his/her intelligence is the starting point of the AI. The technology is intended to imitate human behaviour through mechanical work and understanding in order to apply it independently on this basis, if necessary. For this purpose, huge amounts of data are processed with the aim of identifying patterns from the data, evaluating them and drawing conclusions from them.

"Artificial intelligence - a key to growth and prosperity." This is the title of the 2018 Digital Summit of the Federal Government. Germany is strong in the digital economy and should become a leader in the AI. The strategy is right. However, it is bound by legal guidelines. The General Data Protection Regulation (GDPR) provides a reliable legal framework for innovative technologies and applications, including in the field of AI. It lays down rules on the protection of individuals with regard to the processing of personal data and on the free movement of such data. “The revision of the e-privacy regulation is intended to complement this protection concept.” This is the clear commitment of the Federal Government within its Key Issues Paper as part of Germany’s digital strategy.

In order to make personal data economically usable, the GDPR relies on pseudonymisation. Pseudonymisation serves two functions – one that it is intended to protect personal data, the other to facilitate its economic use at the same time. The core of pseudonymisation is to replace a person's identity data with a specific string, as it is the case with a vehicle registration number. Disclosing a person’s identity from the pseudonym takes place according to fixed rules.

The Focus Group on data protection of the Platform Security, Protection, and Trust For Society and Business has already published a white paper which presents guidelines for the legally secure use of pseudonymisation solutions, taking into account the GDPR.

In 2018, the Focus Group continued its work and presents this working paper. It stipulates requirements for the use of pseudonymisation solutions in compliance with data protection regulations. At the same time, it represents a necessary intermediate step from the whitepaper on the way to a proposal for a code of conduct for pseudonymisation which the Focus Group intends present at the Digital Summit in 2019 and which will help industry to achieve greater investment security.

My sincere thanks is owed to to all the members of the Focus Group for their intensive, constructive and efficient work. Special thanks go to Mr Steffen Weiß for coordinating the Group's work so expertly and prudently.

Cologne, November 2018

Professor Dr. Rolf Schwartmann

Head of the Data Protection Focus Group of the Platform Security, Protection and Trust for Society and Business at the Digital Summit 2018 and member of the Federal Government's Data Ethics Commission

* Available at: https://bit.ly/2FjLnVd.

Table of contents

D.1.11 HMAC

See: Cryptographic checksum.

D.1.12 Homonym error

A homonym error occurs when pseudonymisation procedures that provide linkability falsely lead to the same pseudonyms from different persons.

D.1.13 Identity data

All data relating to a person that make it possible to identify the person in more detail.

D.1.14 Content data

In a data collection, essentially all data that do not belong to the identity data. Nevertheless, a personal reference can be established from content data if they are, for example, unique and this information can be linked to a person.

Remark:

Sometimes there may be overlaps between content data and identity data, e.g. in the data collection for a study to investigate statements about dependency on age or occupation on certain characteristics. In this case, age and occupation would (also) be counted among the content data.

D.1.15 Control number

See: Pseudonym.

D.1.16 Cryptographic hash function

A hash function is a function that assigns a string of any length to a string of fixed length (about 256 bits). A cryptographic hash function also has the property of a one-way function. If, in addition, it is practically impossible to find two different input values that provide the same function value, one speaks of a collision-resistant hash function. Internationally standardized cryptographic hash functions are MD5, SHA256 or SHA-3.

D.1.17 Cryptographic checksum

A bit sequence of fixed length (about 256 bits) that is calculated from a character string of any length using a cryptographic key. If the key is known, the checksum can be used to determine the integrity of the string. Without knowledge of the key, it is impossible to create a valid cryptographic checksum for a character string. An internationally standardized cryptographic checksum is calculated using the HMAC algorithm ((Keyed- Hash Message Authentication Code).

D.1.18 Cryptographic key

A character string that is used to transform a set of data using a cryptographic function (encryption or signature). Depending on the application, the key must be kept secret.

D.1.19 Pseudonym

A character string that replaces the identity data of a person and thus represents this person. The identity data of a pseudonym should, if at all, only be inferred under strictly defined conditions (see Discoverability).

D.1.20 Pseudonymisation list

A list that compares identity data and pseudonyms. A pseudonymisation list can be used to determine a person's pseudonyms directly from an individual's identity data and vice versa to determine an individual's identity data from an individual's pseudonym.

D.1.21 Pseudonymisation stages

If a pseudonym is not created directly from the identity data, but in mutually independent stages, one speaks of pseudonymisation lstages.

Remark:

A pseudonymisation in several stages takes place, for example, with the participation of one or more trust bodies.

D.1.22 Pseudonymisation procedure

A procedure which generates a pseudonym from the identity data of a person.

D.1.23 Record Linkage

In specialist literature, the merging of data records of a pseudonymised data collection on the basis of linkable pseudonyms is referred to as record linkage.

D.1.24 Re-identification

See Discoverability.

D.1.25 Synonym error

Occurs when, in a linkable pseudonymisation procedure, identity data of the same person incorrectly lead to different pseudonyms, although this was not intended.

D.1.26 Linkability of pseudonyms

A pseudonymisation procedure ensures the linkability of pseudonyms if identity data for the same person generally lead to identical or similar pseudonyms. The ps-eudonym, respectively the data records of the person are then "linkable": Identical pseudonyms can usually be used to identify identical persons.

Remarks:

The linking of pseudonymised data with persons without knowledge of the pseudonymisation procedure or the pseudonymisation table is not meant and must be avoided.

In the case of linkable pseudonyms, homonym or synonym errors can nevertheless occur.

D.1.27 Encryption

A method which converts a plaintext into a ciphertext depending on a cryptographic key. The inversion, i.e. to restore the plaintext from the encrypted text, is called decryption.

D.1.28 Trusted third party

A body which is independent of the data controller in terms of space and organisation. The only task of the trusted third party is to support the conversion of identity data into pseudonyms.

Remark:

If necessary, several trust authorities can be involved in a pseudonymisation process, which create the pseudonyms in several pseudonymisation stages.

D.1.29 Allocation table

See pseudonymisation list.

D.2 Measures

D.2.1 General information

In the case of pseudonymisation, basic principles are to be observed which must be observed for every procedure:

a. Knowledge, only if necessary

b. Delete data, whenever possible

c. Avoiding the accumulation of too much knowledge in one place (e.g. plain text data and pseudonymised data about a person)

d. Pseudonyms only if there is a need for them; otherwise anonymization

Depending on the context, different types of pseudonyms can be used:

- Personal-pseudonyms, which are used instead of identity data such as name, ID number or mobile phone number are displayed
- Role-pseudonyms, where one or more persons are assigned to a pseudonym (e.g. IP number)
- Relationship-pseudonyms, in which a person uses a different pseudonym for each (communication) relationship, e.g. different nicknames
- Role-relationship-pseudonyms, which are a combination of the two pseudonym types
- Transaction-pseudonyms, in which a new pseudonym is used for each transaction, which is used, for example, in online banking.

In general, the linkability of personal pseudonyms is considered higher than that of role or relationship pseudonyms. Even less is the linkability of role-relationship pseudonyms and trans-action pseudonyms; in principle they cannot be linked. Basically, the less pseudonymisation can be linked, the greater is the possible anonymity of the data for third parties. A low linkability increases the strength of the pseudonymisation at the same time. In addition, the technical-organisational implementation of a pseudonymisation requires various process steps, which typically are:

D.2.2 Creation of a pseudonym (pseudonymisation of the data record)

Every pseudonymisation begins with the creation of pseudonyms that connect data sets with associated natural persons. The pseudonym may be used to re-identify a data set, must be kept separately and protected by technical and organisational measures.

With the data to be pseudonymised, a distinction is made between identity data of the persons involved and content data. A strict separation between the two types of data is not possible in all cases, so that content data can also contain information about a person (e.g. gender, occupational group and year of birth) and thus an identification of a data subject is possible.

The type of pseudonymisation chosen can have a fundamental influence on the user's scope of action. With a strong pseudonymisation, more critical data processing can usually be sufficiently protected than with a weak pseudonymisation. It is also true in the area of compatible further processing that with stronger pseudonymisation a given compatibility of the intended further processing with the original purpose can be assumed.

There are basically two procedures available for creating a pseudonym: Pseudonymisation lists and pseudonyms by calculation methods.

D.2.2.1 Pseudonymisation lists

A pseudonymisation list is used to assign pseudonyms to identity data using a table. The pseudonyms have no relation to the identity data, neither from functional nor a content perspective.

Example 1:
Pseudonyms are numbered consecutively.

Identity data	Pseudonym
Peter Müller **born** **31.01.1965**	**2022917**
Maria Schulze **born** **03.05.1959**	**2022918**
Max Klein **born 31.10.1967** **in Bornheim**	**2022919**

Example 2:
Pseudonyms are generated randomly or pseudo-randomly.

Identity data	Pseudonym
Peter Müller born 31.01.1965	2184578
Maria Schulze born 03.05.1959	3654425
Max Klein born 31.10.1967 in Bornheim	8745124

Comments:

1. When the pseudonyms are numbered consecutively, it may be possible to draw conclusions about the identity data. For example, if the output data is sorted alphabetically. Or at what time the pseudonyms were created (example: Spanish registration plates provide information about the initial registration of the vehicle).

2. With random pseudonyms, the length of the pseudonyms should not be too short, otherwise collisions and homonym errors can occur. The rule of thumb is, that with n possible pseudonyms, a collision occurs after the square root of n formed pseudonyms with a probability of 50%. So, if the pseudonyms are chosen as ten-digit decimal numbers, after 10000 randomly generated pseudonyms with a probability of 50%, two identical pseudonyms are created (keyword "birthday paradox"[3]).

3. The random function offered by a programming language should not be used as the source of randomness (e.g. the function rand() in the programming language C). For example, the iterated output of a cryptographic hash function can be used as a random source:

A1 = Hash(A0),
Pseudonym1 = Bit 1 to 40 of A1

A2 = Hash(A1),
Pseudonym2 = Bit 1 to 40 of A2

A3 = Hash(A2),
Pseudonym3 = Bit 1 to 40 of A3

A0 is a genuine random value to be chosen by the pseudonymisation authority with an entropy of at least 100 bits. For the selection of the number of bits (here 40) see note 2.

4. If several data suppliers are involved in the pseudonymisation process and if a re-identification of the data supplier on the basis of a pseudonym should be possible, the identity of the data supplier can also be pseudonymised and placed in front of the pseudonyms of the persons.

D.2.2.2 Pseudonyms through calculation methods

Another possibility is to calculate the pseudonyms from identity data using an algorithm.

The transformation process has to consider a state-of-the-art procedure (e.g. the Federal Office for Information Security's guideline TR- 02102-11 or the ENISA guideline on crypto procedures) in order to avoid weak points of an encryption which could lead to the disclosure of a person.

In order not to be able to deduce the identity data (ID) from the pseudonym, the calculation must depend on a certain parameter, a so-called cryptographic key.

Calculation methods can be the following:

Encryption with an encryption method:
Pseudonym = E_K(ID).

Here E_K refers to the encryption with a block cipher algorithm, such as AES, with the key K.

Formation of a cryptographic checksum:
Pseudonym = $HMAC_K$(ID).

HMAC = a Keyed Hash Message Authentication Code, like RFC2104.

Comments:

1. The entropy of K should be at least 100 bits.

2. To calculate the pseudonym, not all identity data need to be used. In general, it is sufficient to make a selection of the identification data so that the person can be identified in the data collection to be pseudonymised. See also section E.2.

3. The entire output of the calculation is not needed to generate the pseudonym. See note 2 from section D.2.2.1.

4. Although a cryptographic hash function is a one-way function, it is not sufficient to calculate the pseudonym exclusively using the hash function, for example via
   - **Pseudonym = Hash(PID)**

If a pseudonym is present, the PID whose hash value gives the pseudonym could be determined by an exhaustive search of all possible values for PID. In Germany, depending on the composition of PID, this search would be limited to a maximum of 80 million hash value calculations.

[3] https://en.wikipedia.org/wiki/Birthday_problem.

5. In a data collection, the identity data may be replaced by several pseudonyms that are calculated from different attributes of the identity data.

Example:

Pseudonym1 =
E_K(health insurance number)

Pseudonym2 =
E_K(Name | Birthday | Place of birth)

Pseudonym3 = E_K(birth name | birthday | place of birth)

6. The generation and administration (e.g. distribution, storage, use, deletion) of secret parameters (cryptographic keys) must be realized by state-of-the-art technical and organisational measures.

7. The security of the chosen pseudonymisation procedure can be increased by defining suitable intervals - depending on time or data volume - in which a secret parameter (cryptographic key) is exchanged. Depending on the type of procedure chosen and the risk for those affected, several pseudonymisation stages can also be built in to exclude detectability (so-called "over-encryption").

D.2.2.3 Multi-stage and mixed pseudonymisation procedures

The security of a pseudonymisation procedure can be increased if the creation of pseudonyms is carried out by several independent bodies. Both pseudonymisation lists and calculation methods can be used.

Example:

1. A, B and C collect data from individuals (A, B and C can be, for example, medical practices that collect patient data).

2. A, B and C form data sets with the help of a calculation method and a cryptographic key K1 (which is available at all data collection points).

3. A, B and C deliver the pseudonymised data records to a trusted third party.

4. V forms new pseudonyms P2 from the obtained pseudonyms P1 using a calculation method and a cryptographic key K2 for the data records and replaces the obtained pseudonyms P1 with the new pseudonyms P2.

5. V forwards the data records with the new pseudonyms P2 to a collector C.

6. C uses the pseudonyms P2 to merge the received data records by means of record linkage.

7. The data are to be evaluated at points X, Y and Z (from different points of view). For this purpose, C filters the data collection and compiles the necessary data records for X, Y and Z from the data collection.

8. From the (partial) data collection for X (and also for Y and Z) the pseudonyms P2 are removed and replaced by new pseudonyms P3, which result from a pseudonymisation list LX, which assigns the pseudonyms P3 to the pseudonyms P2. The pseudo-nymisation lists LX, LY and LZ for X, Y and Z are different and independent of each other.

Remark:

By generating different lists it is ensured that not several data evaluators can merge the data collections made available to them on the basis the pseudonyms contained in these collections.

D.2.2.4 Advantages and disadvantages of different pseudonymisation methods

Method	Advantages	Disadvantages
Assignment tables	**1.** No key management required	**1.** Poor scalability (table can become very large) **2.** Table must be protected permanently **3.** Pseudonymiser needs permanent access to the whole table **4.** Discoverability requires access to the entire table **5.** Linkability requires access to the entire table **6.** Access to the table implies linkability and discoverability (linkability and discoverability are not separately controllable) **7.** Access based on roles requires role-specific table copies
Calculation method	**1.** Good scalability, no table management **2.** Control of knowledge of secret parameters allows access control to calculation rules **3.** Different parameters for pseudonymisation, linkability and discoverability are possible, therefore separately controllable **4.** Only the cryptographic keys need to be securely protected **5.** Role-based access via roll-specific parameters is easily possible **6.** Purpose limitation via technical parameters provides linkability and discoverability based on specific purposes	**1.** Key management required (if necessary further secret or public parameters are needed)

D.2.3 Separate storage of the cryptographic key

D.2.3.1 Access control (authorization concept)

A separate storage of the cryptographic key requires a documented authorization concept. At least two different roles must be defined:

1) The role with access authorization to the key for re-identification;

2) The role with access to the pseudonymised content data.

It is recommended to define the following roles for a pseudonymisation procedure:

1. Provide data

2. Pseudonymise data and re-identify it, if necessary

3. Collect data and merge them using pseudonyms ("record linkage")

4. Evaluate data

It is mandatory that roles 2 and 4 exist separately from each other.

It should be avoided that a person is assigned to multiple roles. This also applies to administrators. Any exceptions must be justified and documented.

Access to a cryptographic key must be restricted to an absolute minimum of trustworthy persons (need-to-know principle).

The possibility of re-identification should not exist in the department of an organisation in which content data belonging to a pseudonym are processed. Any exceptions must be justified and documented.

D.2.3.2 Four-eyes principle

Any access to a cryptographic key for the re-identification of identity data must follow the four-eyes principle. This can be solved technically or organisationally. Furthermore, none of the persons involved should have access rights to both the cryptographic key, the pseudonym and the associated content data. If the four-eyes principle is not possible, at least the access to the cryptographic key must be logged individually.

D.2.4 Documentation of technical and organisational measures for non-assignability

Technical-organisational measures to ensure that a pseudonym cannot be assigned to identity data, for example in the case of missing legitimation, must be documented. This can be done in a pseudonymisation concept. The concept must be integrated into an IT security management system (e.g. ISO/IEC 27001). The IT security management system has to be documented and its effectiveness regularly reviewed.

D.2.5 Rules for disclosure

Since a re-identification of identity data may be possible during pseudonymisation, a planned disclosure of a pseudonym must be regulated. To this end, a documented definition of cases of a desired disclosure is needed. The process of re-identifying the data subject must also be logged. The record must show which persons carried out the re-identification. No conclusions about the identity data on which a pseudonym is based may be drawn from the recording. Therefore, the scope of the logging must be restricted. Log data may only be stored for a limited time.

D.2.6 Loss of purpose for processing

The purposes and duration of the pseudonymisation procedure shall be determined in advance and the measures for the termination of the procedure, including the technical implementation of a data deletion, be documented.

If the purpose for a pseudonymisation no longer applies, e.g. the data is no longer needed, pseudonymised data must be deleted or anonymised in accordance with data protection regulations. Such anonymisation cannot usually be achieved by deleting the pseudonyms, but must take place as an independent procedure for which special requirements apply which cannot be dealt with in detail here. In the case of anonymisation, it must also be checked at regular intervals whether the data can still be classified as anonymous. If a data subject has a right to delete his/her data, this right refers to personal data and pseudonymised data, not to anonymous data. Legal retention periods must be observed.

E. Best practices

E.1 Linkable pseudonymisation methods

A pseudonymisation process provides linkable pseudonyms if identical or similar pseudonyms are generated for persons with the same or similar identity data. In this case, data records can be merged using pseudonyms. Linkable methods are important for long-term studies, for example, or if the data sets come from different sources and are to be merged for one study. The process of merging by means of linkable pseudonyms is referred to in specialist literature as record linkage.

Examples:

1. For studies on the legal probation of offenders, the content data (offence, sentence, age, etc.) are collected in a database. For data protection reasons, the entries may not have any personal reference. Authorities are regularly obliged to delete data on previous convictions of persons after legally prescribed periods of time. However, in order to carry out long-term studies on the recidivism of offenders, the data material could be analysed using linkable pseudonyms.

2. The German epidemiological cancer registries collect pseudonymised data sets on cancer patients in order to investigate the success of different

3. treatment methods. Data suppliers include doctors, hospitals and death

   registers. Some of the data extend over long periods of time and may even originate from different federal states, as the patients may have changed their place of residence. Meaningful studies can only be created on the basis of linkable pseudonyms.

Note:

If there are several pseudonyms in the data collection for a data set (see note 5 in Section D.2.2.2), the data sets can be linked if only one of the pseudonyms matches.

E.2 Selection of identity data

All attributes relating to a person that allow the person to be more closely identified belong to the identity data of the person. These could be for example:

- First name, family name and maiden name
- Gender
- Date and place of birth
- Place of residence and nationality
- Number of siblings
- Occupation or occupational group
- Health insurance or identity card number
- and much more.

E.2.1 Identity data for the calculation of pseudonyms

The identity data of a person can be used, as described in Section D.2.2.2, to calculate the pseudonym to the person.

It has to be taken into account that when using a cryptographic function to calculate the pseudonyms, the same identity data will provide the same pseudonyms, but even minor deviations in the identity data will lead to completely different pseudonyms. Reasons for a change of the pseudonyms can be:

- Writing and typing errors or transposed numbers
- Change of name due to wedding or divorce
- Different spellings of the first name (e.g. Hans/Johannes, Inge/ Ingrid)
- Change of residence
- Change of name of a locality due to a territorial reform
- Ignorance of an attribute (e.g. place of birth)
- and much more.

If the case occurs that a person is assigned different pseudonyms at different times or from different places, one speaks of a synonym error. In this case, the pseudonym can no longer be linked to this person.

The synonym error rate can be reduced by the following measures:

- Omission of an attribute when calculating the pseudonym, for example, only the year of birth is used instead of the complete date of birth.
- Restriction of the name to the initial letter or letters (i.e. the first three)
- Use of a name or phonetic code instead of the name (see for example de.wikipedia.org/wiki/Kölner_Phonetik)
- Use of the municipality code number instead of place of residence or birth
- and much more.

If, on the other hand, different people receive the same pseudonym at different

times or from different places, this is referred to as a homonym error. If the pseudonyms are calculated from the identity data, homonym errors always occur if the identity data from which the pseudonyms are calculated match for both persons.

The homonym error rate can be reduced by the following measures:

- Adding additional attributes for the calculation of the pseudonym, e.g. the complete date of birth can be used instead of only the year of birth.
- Use of long-lasting unique characteristics for calculating pseudonyms, such as the pension insurance or health insurance numbers
- and much more.

Comments:

1. In the case of a high synonym error rate, values are generally underestimated (e.g. the relapse rate in a legal probation study or the mortality rate in a specific treatment method).
2. With a high homonym error rate, values are generally overestimated.
3. A reduction of the synonym error rate usually results in an increase of the homonym error rate - and vice versa.
4. A compromise between synonym error rate and homonym error rate strongly depends on the underlying or expected data collection. Accordingly, the attributes of the identity data to be used for the calculation of the pseudonyms are to be selected.

E.2.1 Identity data in the content data

In pseudonymised data collections, the content data may still contain identification data, provided that this can be of significance for the intended research using the data collection. For example, gender, age, place of residence (as a five-digit postal code) or occupation may be of interest. In certain cases, however, it may be possible to identify individuals solely on the basis of the identity data contained in the content data.

For example, it is conceivable that there is only one floor tiler in the postal code area 65432. This would then undoubtedly be identifiable in the data collection. However, even if there are several floor tilers with the postal code 65432, it would have to be ensured that these do not all have a certain characteristic in common, for example a certain illness, because otherwise one would immediately know from a person of whom one knows that he is a floor tiler by profession and has the postal code 65432 that he suffers from this illness.

For a pseudonymised data collection k-anonymity and l-diversity must therefore be guaranteed.

A data collection offers k-anonymity if the identity data of each individual person contained in it overrides at least k - 1 other persons.

A data collection offers l-diversity if there are at least l different forms of content data for each group of identical identity data contained therein.

k and l are natural numbers.

Comments:

1. Larger values for k and l represent a greater anonymity in this context.
2. k-Anonymity and l-Diversity can be achieved by aggregating the attributes in the identity data.

Examples:

- Instead of "tiler", "craftsman" is indicated as the occupation.
- All postal codes in the data collection that begin with 654 are added together. Instead of 65432, 654xx is then stored in the data collection.

3. k-anonymity and l-diversity shall be established by the pseudonymising body (see Section D.2.3.a). For this purpose, the pseudonymising entity must have access to the attributes of the identity data contained in the content data.

E.3 Involving a trusted third party

The security of pseudonymisation procedures is generally increased if the roles mentioned in Section D.2.3.a are separated organisationally and locally. A trusted third party receives the data collection of the data supplier(s), generates the pseudonyms and forwards them to the data collection entity. The data collection entity then merges the data received using the pseudonyms. The data collection entity then passes them on to the data evaluator(s). In this way, neither the data collection entity nor the data evaluator come into contact with the identity data at any time.

After pseudonymisation at the trusted third party, the trusted party may be obliged to delete the identity data irretrievably if there is no need to re-identify the pseudonyms (see sections D.2.5 and E.4). After completion of the entire procedure, the trusted third party may be obliged to delete the cryptographic keys used.

For the trust third party there is no need to know the content data, but only the identity data in the case of a linkable pseudonymisation procedure. It is therefore advisable to set the content data to a separate transmission path from the data suppliers directly to the data collection entity. The separate transmission path can be of a physical nature; however, the content data can also pass through the trusted third party and be encrypted using an encryption procedure in which only the data collection entity is able to decrypt the data.

E.3 Discoverability of pseudonyms/re-identification

Under certain circumstances, it may be necessary to trace the associated person or his or her identity data from a pseudonym.

In the event that the pseudonym has been created by a calculation procedure from the identification data, it is necessary for discovery that the cryptographic keys used have not been deleted. If the formation of the pseudonyms was based on an encryption process, the pseudonym can be decrypted immediately in order to access the identity data. If the pseudonym was formed by a crypto-graphical checksum, it is not possible to discover the identity data directly. However, if the key K used has not been deleted, the identity data can be determined by a complete exhaustion of all relevant identity data (see comment 4 in Section D.2.2.2).

If the pseudonym was created from the identity data by a pseudonymisation list, it is necessary for the discoverability that the pseudonymisation list used was not deleted.
For multi-stages and mixed processes, all cryptographic keys and pseudonymisation lists used for creation are required for discoverability.
In the example scenario from section D.2.2.3, a re-identification of a pseudonym P3, which is present at the data evaluator X, would be possible as follows:

1. X returns the pseudonym P3 to S

2. S determines the pseudonym P2 from P3 using the list LX.

3. S returns the pseudonym P2 to V

4. V calculates the pseudonym P1 from P2 using the key K2.

5. V returns the pseudonym P1 to an authorized authority that has knowledge of the key K1.

The authorized authority determines the associated identity data from P1 using the key K1.