INTERNATIONAL STANDARD

ISO/IEC 20889

First edition
2018-11

Privacy enhancing data de-identification terminology and classification of techniques

Terminologie et classification des techniques de dé-identification de données pour la protection de la vie privée

Reference number
ISO/IEC 20889:2018(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see http://patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 27, *IT Security techniques*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

It is well-established that major benefits can be derived from processing electronically stored data, including so-called “big data”. However, where this data includes personally identifiable information (PII), as is often the case, processing this data needs to comply with applicable personal data protection principles. The appropriate use of de-identification techniques is an important component of measures to enable the exploitation of the benefits of data processing while maintaining compliance with the relevant ISO/IEC 29100 privacy principles.

The immediate relevance of this document is to personal data protection of natural persons (i.e. PII principals), but the term “data principal”, defined and used in this document, is broader than “PII principal” and, for example, includes organizations and computers.

This document focuses on commonly used techniques for de-identification of structured datasets as well as on datasets containing information about data principals that can be represented logically in the form of a table. In particular, the techniques are applicable to datasets that can be converted to having the form of a table (e.g. data held in key-value databases). It is possible that the techniques described in this document do not apply to more complex datasets, e.g. containing free-form text, images, audio, or video.

The use of de-identification techniques is good practice to mitigate re-identification risk, but does not always guarantee the desired result. This document establishes the notion of a formal privacy measurement model as an approach to the application of data de-identification techniques.

NOTE 1 Annex C clarifies how selected de-identification techniques described in this document are applicable for de-identification of free-form text.

NOTE 2 The application of de-identification techniques can be a privacy risk treatment option arising from a privacy impact assessment, as described in ISO/IEC 29134[32].

The selection of de-identification techniques needs to effectively address the risks of re-identification in a given operational context. There is therefore a need to classify known de-identification techniques using standardized terminology, and to describe their characteristics, including the underlying technologies and the applicability of each technique to the reduction of the risk of re-identification. This is the main goal of this document. The relationship between the terminology used in this document and related terminology used elsewhere (e.g. the notion of anonymization) is described in Annex B. However, the specification of detailed processes for the selection and configuration of de-identification techniques, including assessments of data usefulness and the overall risk from a re-identification attack, is outside the scope of this document.

NOTE 3 Authentication, credential provisioning, and identity proofing are also outside the scope of this document.

De-identification techniques are typically accompanied by technical and other organizational measures to enhance their effectiveness. The use of these measures is also described wherever applicable.

This document provides an overview of core concepts relating to the de-identification of data, and establishes a standard terminology for, and description of, the operation and properties of a range of de-identification techniques. However, it does not specify how these techniques should be managed in a particular use case. It is anticipated that sector-specific framework standards will be developed to provide such guidance.

Privacy enhancing data de-identification terminology and classification of techniques

1 Scope

This document provides a description of privacy-enhancing data de-identification techniques, to be used to describe and design de-identification measures in accordance with the privacy principles in ISO/IEC 29100.

In particular, this document specifies terminology, a classification of de-identification techniques according to their characteristics, and their applicability for reducing the risk of re-identification.

This document is applicable to all types and sizes of organizations, including public and private companies, government entities, and not-for-profit organizations, that are PII controllers or PII processors acting on a controller's behalf, implementing data de-identification processes for privacy enhancing purposes.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 27000, *Information technology — Security techniques — Information security management systems — Overview and vocabulary*

ISO/IEC 29100, *Information technology — Security techniques — Privacy framework*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 27000, ISO/IEC 29100 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1
aggregated data
data representing a group of *data principals* (3.4), such as a collection of statistical properties of that group

3.2
attribute
inherent characteristic

[SOURCE: ISO 9241-302:2008, 3.4.2]

3.3
formal privacy measurement model
approach to the application of data *de-identification techniques* (3.7) that enables the calculation of *re-identification risk* (3.33)

3.4
data principal
entity to which data relates

Note 1 to entry: The term “data principal” is broader than “PII principal” (or “data subject” as used elsewhere), and is able to denote any entity such as a person, an organization, a device, or a software application.

3.5
dataset
collection of data

[SOURCE: ISO 19115-1:2014, 4.3, modified — The word “identifiable” has been deleted in the definition.]

3.6
de-identification process
process of removing the association between a set of *identifying attributes* (3.14) and the *data principal* (3.4)

3.7
de-identification technique
method for transforming a *dataset* (3.5) with the objective of reducing the extent to which information is able to be associated with individual *data principals* (3.4)

3.8
de-identified dataset
dataset (3.5) resulting from the application of a *de-identification process* (3.6)

3.9
differential privacy
formal privacy measurement model (3.3) that ensures that the probability distribution of the output from a statistical analysis differs by at most a specified value, whether or not any particular *data principal* (3.4) is represented in the input dataset

Note 1 to entry: More specifically, differential privacy provides:

a) a mathematical definition of privacy which posits that, for the outcome of any statistical analysis to be considered privacy-preserving, the analysis results from the original dataset are indistinguishable from those obtained if any data principal is added to or removed from the dataset; and

b) a measure of privacy that enables monitoring of cumulative privacy loss and setting of an upper bound (or “budget”) for loss limit. A formal definition is as follows. Let ε be a positive real number, and M be a randomized algorithm that takes a dataset as input. The algorithm M is said to be ε-differentially private if for all datasets D_1 and D_2 that differ in a single element (i.e. the data for one data principal), and all subsets S of the range of M, $\Pr\left[M\left(D_1\right)\in S\right]\le e^{\varepsilon}\Pr\left[M\left(D_2\right)\in S\right]$, where the probability is taken over the randomness used by the algorithm.

3.10
direct identifier
attribute (3.2) that alone enables unique identification of a *data principal* (3.4) within a specific operational context

Note 1 to entry: Here and throughout, the operational context includes the information that the entity processing (e.g. de-identifying) the data possesses, together with information that third parties and potential attackers can possess or that is in the public domain.

3.11
equivalence class
set of *records* (3.30) in a *dataset* (3.5) that have the same values for a specified subset of *attributes* (3.2)

3.12
generalization
category of *de-identification techniques* (3.7) that reduce the granularity of information contained in a selected *attribute* (3.2) or in a set of related *attributes* (3.2) in a *dataset* (3.5)

3.13
identifier
set of *attributes* (3.2) in a *dataset* (3.5) that enables unique identification of a *data principal* (3.4) within a specific operational context

Note 1 to entry: See Annex B for a discussion of how this definition relates to those given in other standards.

3.14
identifying attribute
attribute (3.2) in a *dataset* (3.5) that is able to contribute to uniquely identifying a *data principal* (3.4) within a specific operational context

3.15
identity disclosure
re-identification (3.31) event in which an entity correctly assigns an identity to a *data principal* (3.4)

3.16
indirect identifier
attribute (3.2) that, together with other *attributes* (3.2) that can be in the *dataset* (3.5) or external to it, enables unique identification of a *data principal* (3.4) within a specific operational context

3.17
inference
act of deducing otherwise unknown information with non-negligible probability, using the values of one or more *attributes* (3.2) or by correlating external data sources

Note 1 to entry: The deduced information can be the value of one or more attributes of a data principal, the presence or absence of a data principal in a dataset, or the value of one or more statistics for a population or segment of a population.

3.18
***K*-anonymity**
formal privacy measurement model (3.3) that ensures that for each *identifier* (3.13) in a *dataset* (3.5) there is a corresponding *equivalence class* (3.11) containing at least *K records* (3.30)

3.19
***L*-diversity**
formal privacy measurement model (3.3) that ensures that for a selected *attribute* (3.2) each *equivalence class* (3.11) has at least *L* well-represented values

Note 1 to entry: *L*-diversity is a property of a dataset that gives a guaranteed lower bound, *L*, on the diversity of values shared by an equivalence class for a selected attribute.

3.20
linkability
property for a *dataset* (3.5) that it is possible to associate (by linking) a *record* (3.30) concerning a *data principal* (3.4) with a *record* (3.30) concerning the same *data* principal in a separate dataset

3.21
linking
act of associating a *record* (3.30) concerning a *data principal* (3.4) with a *record* (3.30) concerning the same data principal in a separate *dataset* (3.5)

3.22
macrodata
dataset (3.5) comprised of *aggregated data* (3.1)

3.23
microdata
dataset (3.5) comprised of *records* (3.30) related to individual *data principals* (3.4)

3.24
noise addition
de-identification technique (3.7) that modifies a *dataset* (3.5) by adding random values to the values of a selected *attribute* (3.2)

3.25
permutation
de-identification technique (3.7) for reordering the values of a selected *attribute* (3.2) across the *records* (3.30) in a *dataset* (3.5) without modifying these values

3.26
pseudonym
unique identifier created for a *data principal* (3.4) to replace the commonly used *identifier* (3.13) [or *identifiers* (3.13)] for that *data principal* (3.4)

Note 1 to entry: A pseudonym is sometimes also known as an alias.

3.27
pseudonymization
de-identification technique (3.7) that replaces an identifier (or identifiers) for a *data principal* (3.4) with a *pseudonym* (3.26) in order to hide the identity of that data principal

3.28
quasi-identifier
attribute (3.2) in a *dataset* (3.5) that, when considered in conjunction with other attributes in the dataset, *singles out* (3.35) a *data principal* (3.4)

3.29
randomization technique
de-identification technique (3.7) in which the values of *attributes* (3.2) are modified so that their new values differ from their true values in a random way

3.30
record
set of *attributes* (3.2) concerning a single *data principal* (3.4)

3.31
re-identification
process of associating data in a de-identified *dataset* (3.5) with the original *data principal* (3.4)

Note 1 to entry: A process that establishes the presence of a particular data principal in a dataset is included in this definition.

3.32
re-identification attack
action performed on de-identified data by an attacker with the purpose of *re-identification* (3.31)

3.33
re-identification risk
risk of a successful *re-identification attack* (3.32)

3.34
sensitive attribute
attribute (3.2) in a *dataset* (3.5) that, depending on the application context, merits specific, high-level protection against potential *re-identification attacks* (3.32) enabling disclosure of its values, its existence, or association with any of the *data principals* (3.4)

Note 1 to entry: Designating an attribute as sensitive depends on the application context, and such a designation is an input to the design of the *de-identification process* (3.6) in a specific use case.

3.35
single out
isolate *records* (3.30) belonging to a *data principal* (3.4) in the *dataset* (3.5) by observing a set of characteristics known to uniquely identify this data principal

3.36
***T*-closeness**
formal privacy measurement model (3.3) that ensures that the distance between the distribution of a selected *attribute* (3.2) in an *equivalence class* (3.11) and the distribution of this *attribute* (3.2) in the entire table is no more than a threshold *T*

Note 1 to entry: A table is said to have *T*-closeness with respect to a selected attribute if all equivalence classes containing this attribute have *T*-closeness.

3.37
truthful data
factual data that has not been accidentally or deliberately distorted

Note 1 to entry: Reducing granularity maintains data truthfulness.

3.38
unique identifier
attribute (3.2) in a *dataset* (3.5) that alone singles out a *data principal* (3.4) in the dataset

3.39
usefulness
degree of suitability of the type and format of information in a *dataset* (3.5) for application to a specific purpose

Note 1 to entry: The term *utility* is sometimes used with a similar meaning.

4 Symbols and abbreviated terms

ICO	Information Commissioner's Office (UK)
PII	Personally Identifiable Information
PPDM	Privacy-Preserving Data Mining
PPDP	Privacy-Preserving Data Publishing
SDC	Statistical Disclosure Control
SDL	Statistical Disclosure Limitation
ε	privacy budget
C	privacy cost
D_1, D_2	datasets
k	number of records or a parameter (percentage) used in the threshold rule
K	parameter used in the *K*-anonymity model
L	parameter used in the *L*-diversity model
M	randomized algorithm
N	number of queries posed

n	threshold value or number of queries/values (depending on context)
p	parameter (percentage) used in the ambiguity rule
Pr	probability function
q	parameter (percentage) used in the ambiguity rule
S	sensitivity of a query
T	parameter used in the T-closenesss model
$\in$	is an element of (a set)

5 Overview

The goal of this document is to provide organizations that are implementing privacy and security measures with information to support the selection and design of appropriate de-identification techniques in the context of their organization. After introducing the terminology used in this document in Clause 6, the threat of re-identification is described in Clause 7; the risk of re-identification attacks must be assessed as part of the process of selecting appropriate de-identification techniques. The degree to which usefulness of data after de-identification is retained is the subject of Clause 8.

De-identification techniques are classified in Clause 9 according to their underlying technologies. Each technique is further assessed in terms of its effectiveness against re-identification as well as its usefulness in a variety of use cases. The organization's objectives provide an essential context for the choice of de-identification techniques and other supporting measures. While definition of security technologies is outside the scope of this document, their use as part of de-identification techniques is described wherever applicable. The use of technical and other organizational measures to enhance the effectiveness of de-identification techniques is also described wherever applicable.

NOTE A de-identification process can be deemed to be successful if the output is de-identified to the required level. This level can differ between two processes depending on the use case (e.g. the dataset dimension, the type and nature of the data, the type and nature of the data principal, the motivation of the attacker, and the availability of other sources of information).

Formal privacy measurement models provide a means of assessing the effectiveness of de-identification and are described in Clause 10. This is followed in Clause 11 by a discussion of principles for the use of de-identification techniques. The main body of the document concludes in Clause 12 with a review of additional technical or organizational measures that can be applied for de-identification.

6 Technical model and terminology

De-identification refers to a process that removes the association between a set of data attributes and the data principal which they concern. A de-identification technique is a method for transforming a dataset with the objective of reducing the extent to which data can be associated with specific data principals.

Unless otherwise stated, the focus of this document is on de-identification of datasets that are represented as a collection of records, where each record is comprised of a set of attributes. Each record carries information about a data principal and each attribute carries a known meaning.

Regardless of its physical or logical arrangement, any dataset that fits these characteristics is capable of being viewed as a table. That is, it can be represented as a table with rows and columns, where each data principal is represented by a single row and each column represents an attribute with a known meaning. This includes key-value databases, i.e. repositories for data consisting of pairs made up of a key and a value. The assumption that each data principal is only associated with a single row is made for the purposes of this document in giving a conceptual description of de-identification techniques and formal privacy measurement models. However, in practice, data principals are sometimes represented by multiple rows (e.g. because of the formal privacy measurement model considered), or in aggregated form with attributes represented by both rows and columns.

NOTE 1 If, in practice, multiple rows exist corresponding to a single data principal, then, if possible, for the purposes of applying the concepts described in this document they need to be merged.

Some datasets are not organized as a collection of records each carrying information about a single data principal. Examples of such datasets are free-form text files, log files of events recorded in order of their occurrence, and logical graphs that indirectly contain information about data principals. Such datasets can still contain fields carrying personal or sensitive characteristics relating to potential data principals.

Annex C describes a number of existing approaches to the de-identification of free-form text, and clarifies how some of the de-identification techniques described in this document are able to be used for the de-identification of free-form text. Transactional/longitudinal data can be reorganized into a binary view with each data principal represented by a single row (see Annex D).

The classification of attributes used in this document, as summarized below, reflects the difference between singling out a data principal in a dataset and identifying a data principal within a specific operational context, where the operational context includes information possessed by the entity processing (e.g. de-identifying) the data, together with information that third parties and potential attackers can possess:

NOTE 2 Singling out a data principal in the dataset is a necessary precursor to singling out the data principal in the population; re-identification occurs if the data principal can be singled out in the population, but not necessarily if the data principal is only singled out in the dataset (which can be a sample).

— Identifier (see 3.12): a set of attributes in a dataset that enables unique identification of a data principal within a specific operational context;

— Local identifier: a set of attributes in a dataset that together single out a data principal in the dataset;

— Direct identifier (see 3.9): an attribute in a dataset that alone enables unique identification of a data principal within a specific operational context;

— Unique identifier (see 3.37): an attribute in a dataset that alone singles out a data principal in the dataset;

NOTE 3 Pseudonyms are examples of unique identifiers.

— Indirect identifier (see 3.15): an attribute in a dataset that, when considered in conjunction with other attributes that can be in the dataset or external to it, enables unique identification of a data principal within a specific operational context;

NOTE 4 A related term sometimes used to describe “indirect identifier” is “key attribute”.

NOTE 5 In the context of PII, examples of indirect identifiers include: birth date, postal area code (ZIP code), and sex.

NOTE 6 There is the potential for a large proportion of the attributes in some datasets to qualify as indirect identifiers. This can be considered during the selection of de-identification techniques.

— Quasi-identifier (see 3.27): an attribute in a dataset that, when considered in conjunction with other attributes in the dataset, singles out a data principal;

Sensitive attributes (see 3.33) require specific, high-level protection against potential re-identification attacks enabling disclosure of their values, existence, or association with any of the data principals. Designating an attribute as sensitive depends on the application context.

NOTE 7 In some contexts (such as in specific jurisdictions), attributes are deemed sensitive depending on the nature of the PII, and can, for example, include racial or ethnic origin, political opinions, religious or other beliefs, personal data on health, sex life or criminal convictions.

Figure 1 depicts the relationships between the various types of identifier described above. Any attribute that is equal to or part of one of the types of identifier defined above is deemed to be an identifying attribute. In Figure 1, "Dataset" refers to the dataset to which de-identification techniques are to be applied, and "context" refers to information derived from the operational context of the dataset.

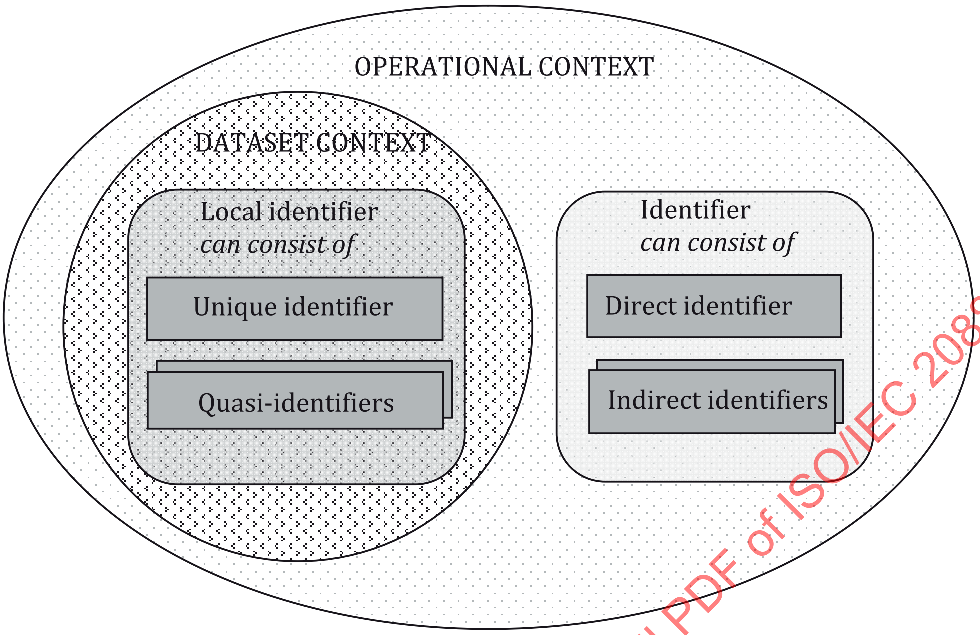

Figure 1 — Types of identifier

7 Re-identification

7.1 General

A de-identification technique is designed to reduce the risk of re-identification by generating data that is less vulnerable to known re-identification attacks. Typically, a de-identification technique alone cannot provide quantifiable guarantees against re-identification attacks. To achieve this, a range of formal privacy measurement models have been devised, enabling the calculation of the risk of re-identification – core models are described in Clause 10. General considerations regarding the effectiveness of de-identification techniques are described as known at the time of publication. The vulnerability of a specific data processing system to re-identification attacks can only be assessed in the context of the organization, and is outside the scope of this document.

Re-identification events in which an entity correctly assigns an identity to a data principal, are known as "identity disclosures". Re-identification events in which an entity correctly assigns data from a de-identified set to a data principal are known as "attribute disclosures". The two disclosure types can help enable each other. For example, using additional datasets, identity disclosure can help enable attribute disclosures, while attribute disclosure for an unknown data principal can contribute to its identity disclosure.

7.2 Re-identification attacks

A re-identification attack is an action performed on de-identified data by an attacker with the purpose of re-identification.

Typically, a re-identification attack involves the creation of an "observation" dataset representing some or all of the data principals from the original dataset. Note that it is possible that the resulting information about the data principals is not identical to or consistent with the original data as a result of modifications to the original values of data attributes in the course of data de-identification, its subsequent re-identification, or a combination of both.

Exact disclosure occurs when an attacker determines the exact value of an attribute for a data principal. Statistical disclosure occurs when aggregated data enables an attacker to obtain a better estimate of an attribute value than is possible without it. More generally, deterministic re-identification happens when an attacker correctly associates the data in a de-identified dataset with the original data principal without using statistics in the process of re-identification.

In this document, an attacker is an entity (either a person or an automated tool) that has access to the de-identified data, in the form dictated by the design of the de-identification technique (e.g. by retrieving the de-identified dataset or through de-identified responses to data queries), as well as access to any additional reasonably available data external to the de-identified data. Given this definition, the cost of implementation, the amount of time required, the technology at the time of the processing, and the technological developments available to the attacker are important considerations in the process of selecting de-identification techniques.

To separate the discussion of de-identification techniques from the security measures implemented by the data processing system, this document focuses on de-identification techniques and formal privacy measurement models and their effectiveness against re-identification performed by a resourceful attacker without breaching technical or other organizational measures.

While attackers differ in their motivations for the act of data re-identification (which can include a proof of concept demonstration or receipt of monetary benefits), known re-identification attacks can be classified according to their goals, as follows:

— Re-identify a record belonging to a specific data principal, possibly using pre-existing knowledge.

NOTE 1 Such an attack is sometimes referred to as a “Prosecutor attack” (see for example, Lan et al.[38] and Wjst[58]) or a “Prosecutor risk” (see, for example, El Emam and Dankar[13]).

— Re-identify the data principal of a specific record, possibly using pre-existing knowledge.

NOTE 2 Such an attack is sometimes referred to as a “Journalist attack” or a “Journalist risk” (see, for example, El Emam and Dankar[13]).

— Re-identify as many records with their corresponding data principals as possible, possibly using pre-existing knowledge.

NOTE 3 Such an attack is sometimes referred to as a “Marketer attack” or “Marketer risk” (see, for example, Dankar and El Emam[6]).

— Establish the presence of a specific data principal in a dataset.

NOTE 4 Such an attack is sometimes referred to as a “(In-)distinguishability attack” or a “data membership attack”) – see ISO 25237:2017, B.2[27].

— Deduce a sensitive attribute associated with a group of other attributes.

NOTE 5 Such an attack is sometimes referred to as a “inference attack” (see, for example, Sweeney[54]).

Despite the differences in goals of re-identification attacks, an attacker typically employs a combination of re-identification approaches. As a result, in a given operational context all applicable re-identification approaches need to be considered.

Known approaches used in re-identification attacks include, but are not limited to:

— singling out: isolating some or all records belonging to a data principal in the dataset by observing a set of characteristics known to uniquely identify this data principal;

— linking: associating records concerning the same data principal or a group of data principals across separate datasets;

— inference: deducing with non-negligible probability the value of an attribute from the values of a set of other attributes.

NOTE 6 An inference of the type described above does not always equate to an attack, since legitimate data analyses can attempt to do so; for example, in a clinical trial one can attempt to establish whether groups of people with a certain set of symptoms can be more prone to a particular condition.

8 Usefulness of de-identified data

De-identified data can be collected, processed, stored, or shared for a wide range of applications and purposes. Each application requires the de-identified data to possess specific properties in order to accomplish its purpose. It is therefore necessary to preserve these properties after de-identification. For example:

— testing of software applications requires data that pertains to or emulates certain characteristics of the anticipated real data in order to achieve behaviour under the test as close as possible to the conditions that will apply during use of the application;

— statistical reporting includes collecting data at the level of individual data principals and generating statistical reports for a population on certain characteristics or events;

— publishing of data for research purposes [also known as Privacy-Preserving Data Publishing (PPDP)] often involves sharing sensitive data at the level of individual data principals;

— performing data analytics on behalf of another party [also known as Privacy-Preserving Data Mining (PPDM)] requires the transfer of data at the level of individual data principals as well as of statistical data;

— accessing and processing of sensitive, truthful, unencrypted data at the level of individual data principals by authorized internal parties in data centres;

— linking data to its corresponding data principal in certain cases by specially appointed parties.

To ensure that the de-identified data is able to be used for its intended purposes, the desired usefulness of de-identified data needs to be determined before selecting the techniques appropriate for each case.

NOTE The de-identification techniques applied to individual attributes in a dataset can differ. For example, in data analytics, modification of the attributes subject to research is typically minimized, whereas other attributes can be more significantly modified to achieve an appropriate level of de-identification,

9 De-identification techniques

9.1 Statistical tools

9.1.1 General

Subclause 9.1 describes methods of a statistical nature that change the overall structure of the data. Such methods are commonly used to either de-identify datasets or to enhance the effectiveness of de-identification techniques.

NOTE The relationship between the terms used in this document and those employed in the field of Statistical Disclosure Control is discussed in B.4.

9.1.2 Sampling

Data sampling is a statistical analysis technique that selects a representative subset of a larger dataset in order to analyse and recognize patterns in the original dataset. To reduce the risk of re-identification, sampling is performed on data principals.

Performing a random sampling adds uncertainty about the dataset. For example, an attacker, by merely matching attributes of a certain record from the sample with external information, cannot be sure that the record corresponds to the specific data principal since there is no certainty that the data principal is present in the sample dataset. More generally, applying generalization or randomization techniques on a sample, rather than on a whole population, can increase the effectiveness of these de-identification techniques.

The methods for drawing samples from data vary broadly and their selection depends on the dataset and the anticipated use cases. An example of a common algorithm is a simple probability sampling where random numbers are used to select the records in a dataset ensuring that there is no correlation among the records in the resultant sample.

The output of sampling used in this way is microdata.

9.1.3 Aggregation

Aggregation involves the combination of related attributes, or of attribute values, to provide information at a broader level than at which detailed observations are taken.

- Aggregation of attribute values includes the set of broadly used statistical functions that, when applied to an attribute in microdata, produce results that represent all the records in the original dataset. The resulting statistical aggregated values are intended to be useful for the purposes of reporting on or analysis of the data as a whole without revealing any individual records.

- Aggregation of related attributes includes attributes within a common branch of a hierarchy, or statistical functions that produce results that represent all the attribute values of the related attributes in any individual record. The resulting aggregated attributes remain within their original records.

If done effectively, aggregation reduces the ability of an attacker to single out, link, or deduce the value of an attribute from the values of one or more of the other attributes in the same data record.

The output of aggregation is macrodata (i.e. data that is aggregated in frequency or magnitude).

9.2 Cryptographic tools

9.2.1 General

Cryptographic tools can be used to implement security measures that enhance the effectiveness of de-identification techniques, as well as acting as a part of de-identification techniques themselves (see, for example, 9.4.3.3). Subclause 9.2 only describes tools that are usable as de-identification techniques in their own right. When used as de-identification techniques the output is microdata.

NOTE 1 In general, using cryptographic techniques makes the implementation of controlled re-identification simpler (see 9.4), but using cryptography for de-identification can be more computationally demanding than using non-cryptographic techniques.

NOTE 2 Whenever cryptographic keys are used, great care needs to be taken to safeguard them from unauthorized access. General guidance on key management principles is provided in ISO/IEC 11770-1[63].

9.2.2 Deterministic encryption

Deterministic encryption is a form of non-randomized encryption, as standardized in ISO/IEC 18033 (all parts)[24]. When employed as part of a de-identification technique, deterministic encryption can be used to replace any identifying or sensitive attribute within a data record with an encrypted value.

The property of deterministic encryption that enables the usefulness of the de-identified data is that two equal values encrypted under the same private key produce two equal ciphertexts. Deterministic encryption does not reduce the truthfulness of the data.

Deterministic encryption produces microdata that is capable of being searched for exact matches, linked and analysed (e.g. discrete analysis of frequencies or distributions). Deterministic encryption retains data usefulness for limited statistical processing and limited privacy-preserving data mining. Analytical operations on deterministically encrypted data are limited to equality checking.

Full re-identification of deterministically encrypted data is only implementable for a party in possession of the appropriate cryptographic key. Re-identification attacks on deterministically encrypted attributes are limited to the extent that the attributes can be analysed without having access to the cryptographic key. Linkability attacks are possible only for attributes deterministically encrypted using the same key. The success of these attacks depends on the choice of parameters for the encryption algorithm.

NOTE 1 Controlled re-identification of deterministically encrypted data is possible for the party controlling the private key.

NOTE 2 Data encryption techniques are standardized in ISO/IEC 18033 (all parts)[24]. Deterministic encryption techniques are not explicitly standardized in ISO/IEC 18033 (all parts), but certain of the symmetric encryption techniques standardized there can be used to realize deterministic encryption.

9.2.3 Order-preserving encryption

Order-preserving encryption is a form of non-randomized symmetric encryption. When employed as part of a de-identification technique, order-preserving encryption can be used to replace any identifying or sensitive attribute within a data record with an encrypted value.

The property of order-preserving encryption that enables the usefulness of the de-identified data is that two values encrypted with the same private key retain the ordering of values in the ciphertexts. For example, if two values have a fixed ordering, then the same ordering holds in the encrypted values. Order-preserving encryption provides a higher level of usefulness for the de-identified data than deterministic encryption. Order-preserving encryption does not reduce the truthfulness of the data.

Order-preserving encryption produces microdata that is capable of being searched for range matches and analysed (e.g. continuous analysis of frequencies or distributions). Re-identification attacks on deterministically encrypted attributes are limited to the extent that the attributes can be analysed without having access to the cryptographic key. Order-preserving encryption retains data usefulness for limited statistical processing and limited privacy-preserving data mining and also secure outsourced storage and processing of data. Analytical operations on data encrypted using an order-preserving technique are limited to checking equality and order relations (e.g. greater-than and less-than).

Full re-identification of data encrypted using an order-preserving technique is only possible for a party in possession of the appropriate cryptographic key. The success of a linkability attack depends on the choice of parameters for the order-preserving encryption scheme used for the attribute.

NOTE 1 Controlled re-identification of data encrypted in an order-preserving way is possible for the party controlling the private key.

NOTE 2 Order-preserving encryption techniques are not currently standardized in any parts of ISO/IEC 18033[24].

9.2.4 Format-preserving encryption

Format-preserving encryption is designed for data that is not necessarily binary. In particular, given any finite set of symbols, like the decimal numerals, a method for format-preserving encryption transforms data that is formatted as a sequence of the symbols in such a way that the encrypted form of the data has the same format, including the length, as the original data. Thus, for example, a format-preserving-encrypted 9-digit social security number is a sequence of nine decimal digits.

Format-preserving encryption facilitates the de-identification or pseudonymization of sensitive information, as well as the retrofitting of encryption technology to legacy applications where a conventional encryption mode is not feasible.

9.2.5 Homomorphic encryption

Homomorphic encryption is a form of randomized encryption. When employed as part of a de-identification technique, homomorphic encryption is able to be used to replace any identifying or sensitive attribute within a data record with an encrypted value.

The property of homomorphic encryption that enables the usefulness of the de-identified data is that two values encrypted with the same public key can be combined with the homomorphic operator of the cryptographic scheme to produce a new ciphertext representing the result of the operation on the de-identified values. Homomorphic encryption typically has a significantly lower performance and higher storage cost than deterministic encryption. However, it enables limited processing of the de-identified data without re-identifying the data. Homomorphic encryption does not reduce the truthfulness of the data.

Homomorphic encryption produces microdata that can be processed within the limits of the homomorphic operations, without requiring access to the private key necessary to decrypt the data. This enables a third party data handler to process the data in a way such that the result of the processing is also encrypted under the private key. While some homomorphic encryption schemes support a single secure operation, partially and fully homomorphic encryption schemes can provide more than one secure operation (e.g. secure addition and multiplication of encrypted values without decrypting the values). Homomorphic encryption schemes are semantically secure, making re-identification attacks infeasible without access to the appropriate private key.

Homomorphic encryption schemes retain data usefulness for limited processing and privacy-preserving data mining and also secure outsourced storage and processing of data. Full re-identification of homomorphically encrypted data is only possible for a party in possession of the private key matching the public key used for encrypting the data.

NOTE 1 Controlled re-identification of homomorphically encrypted data is possible for the party controlling the private key.

NOTE 2 Homomorphic encryption techniques are standardized in ISO/IEC 18033-6[24].

NOTE 3 Homomorphic encryption and secure multi-party computation are related cryptographic concepts, and secure multi-party computation schemes are capable of being used to help realise certain types of homomorphic encryption. Secure multi-party computation can therefore be used as part of a data de-identification process.

9.2.6 Homomorphic secret sharing

Homomorphic secret sharing enables a secret to be divided into “shares”, specified subsets of which are usable to reconstruct the secret, such that if the same mathematical operation is performed on all the shares used to reconstruct the secret then the result is the effect of performing this mathematical operation on the original secret. When employed as part of a de-identification technique, homomorphic secret sharing can be used to replace any identifying or sensitive attribute within a data record with two or more shares produced by a message sharing algorithm. These shares can then be distributed to two or more share-holders, the number of which is determined by the instantiation of the secret sharing scheme.

The property of homomorphic secret sharing that enables the usefulness of the de-identified data is that two values secret-shared among the same share-holders can be combined with the homomorphic operation of the cryptographic scheme to produce new shares representing the result of the operation on the original attribute. Furthermore, homomorphic secret sharing can be combined with secure multi-party computation to perform any secure operation on the de-identified data. Homomorphic secret sharing does not reduce the truthfulness of the data.

Homomorphic secret sharing has a relatively low performance overhead for computation, but has an additional overhead incurred with exchanging shares with the share-holders. The storage overhead of secret-shared data is limited. The homomorphic processing of data de-identified with homomorphic secret sharing is limited, but has a negligible overhead. The processing of data de-identified with homomorphic secret sharing using secure multi-party computation is flexible, but can come at a significant performance cost, depending on the scheme used.

Homomorphic secret sharing produces distributed instances of microdata that are able to be processed within the limits of the homomorphic operation or using secure multi-party computation. Homomorphic encryption schemes are randomized, making re-identification attacks infeasible without control over all share-holders.

NOTE 1 Controlled re-identification of homomorphically secret shared data is possible if the necessary number of shareholders holding shares of the de-identified data agree to the re-identification.

Homomorphic secret sharing retains full usefulness of de-identified data when both the homomorphic operation and secure multi-party computation are used. This makes the de-identification technology potentially useful for joint privacy-preserving statistical analysis and data mining and also secure data sharing and outsourced storage and processing of data. Full re-identification of homomorphically secret-shared data is only possible for a party having control over the number of share-holders determined by the instantiation of the secret-sharing scheme.

NOTE 2 Secret sharing techniques are standardized in ISO/IEC 19592 (all parts)[25]. The use of secure multi-party computation for processing of de-identified personal data is described in ISO/IEC 29101[31].

9.3 Suppression techniques

9.3.1 General

Suppression techniques involve removing selected attributes across all records (e.g. masking), selected attribute values (e.g. local suppression), or selected records from a dataset (e.g. record suppression). Suppression techniques are mostly applicable to categorical data (see 11.6).

Suppression techniques are relatively easy to implement and they preserve data truthfulness. Suppression techniques result in microdata. The drawbacks include the inevitable loss of information and the need to be combined with additional de-identification techniques to achieve robust de-identification results.

9.3.2 Masking

The term “masking” refers to a de-identification technique that involves removing all direct identifiers from the dataset, and potentially stripping out some or all of the additional remaining identifying attributes for all records in the dataset. Removing a portion of a direct identifier so that it is no longer a unique identifier is also considered to be a masking technique.

NOTE 1 However, removing a portion of a direct identifier so that it is no longer a direct identifier but still a unique identifier is considered to be a pseudonymization technique (see 9.4).

NOTE 2 For further information see, for example, Appendix 2 of the ICO anonymisation code of practice[20].

The output of masking is microdata.

After masking has been performed, typically additional de-identification techniques are applied to the dataset, such as those described in 9.4, 9.5, 9.6 and 9.7.

In systems in which masking is the only de-identification techniques being used, protection of the de-identified data is typically reinforced by technical and other organizational measures. Masking is also known by a number of different names that are listed below for reference:

- partial data removal: this term refers to the case where not all identifying attributes, and possibly not even all local identifiers (see Clause 6), are removed during the process of masking;
- data quarantining: this term refers to the case where masking needs to be accompanied by stringent security measures to ensure only authorized access to the dataset, such as access control or contractual terms;

- data limitation: this term refers to the case where data suppression is performed at the time of collection in the context of the specified purpose;
- complete data removal: this term refers to the case where all identifying attributes and all local identifiers (see Clause 6) are removed during the process of masking.

9.3.3 Local suppression

The term "local suppression" refers to a de-identification technique that involves removing specific values of attributes from selected records that, in combination with other identifying attributes, can identify the data principal. Typically, local suppression is applied to remove rare values (or rare combinations of values) of indirect identifiers that still appear after generalization (9.6) has been performed. Local suppression is most often applied to categorical values, while generalization is typically applied to numerical values with the common goal of increasing the number of records sharing the values of their identifying attributes.

9.3.4 Record suppression

The term "record suppression" refers to a de-identification technique that involves removing an entire record or records from a dataset. Typical candidates for removal are records that contain rare combinations of attributes (e.g. outliers).

9.4 Pseudonymization techniques

9.4.1 General

The term "pseudonymization" refers to a category of de-identification techniques that involve replacing a data principal's identifier (or identifiers) with indirect identifiers specifically created for each data principal.

As such, pseudonymization is a technique that enables linking of associated records from different datasets without revealing the identities of the data principals.

Pseudonymization used alone does not reduce the risk that an individual data principal can be singled out.

NOTE For further discussion of these issues, see the Article 29 opinion on anonymization techniques[1].

The output of pseudonymization is microdata.

A pseudonymization process generates supplementary information that can include identifiers removed from the original dataset, pseudonym assignment tables, or cryptographic keys, as described in 9.4.2 and 9.4.3. Such information can be used in the process of a controlled re-identification. To protect this information from re-identification attacks, appropriate technical and other organizational measures need to be applied to such supplementary information in accordance with the organization's objectives and re-identification risk assessment.

9.4.2 Selection of attributes

In order to achieve the desired level of de-identification, it is important to correctly select the subset of identifying attributes to be replaced with the pseudonym. Pseudonymization involves replacing all direct identifiers and potentially some additional or all remaining identifying attributes with a pseudonym. Examples of guidance on attributes to be replaced by pseudonyms are in ISO/IEC 29100[29] and ISO 25237[27]. However, the selection of attributes is specific to each use case and needs to be performed in accordance with the organization's objectives and re-identification risk assessment.

9.4.3 Creation of pseudonyms

9.4.3.1 General

Various techniques can be used to create pseudonyms. The choice of technique is based on factors such as the costs of creating the pseudonyms, the collision-resistance factor of a hash function (i.e. the probability of two inputs hashing to the same output), and the means by which the data principal can be re-identified for the purposes of a controlled re-identification.

9.4.3.2 Pseudonyms independent of identifying attributes

The pseudonym values can be independent of the replaced attributes' original values. Such methods include generation of random values.

When pseudonyms are generated independently of the attributes, a table containing the mappings (or assignments) of the original identifier(s) to the corresponding pseudonym can be created. For security reasons, appropriate technical and organizational security measures need to be applied to limit and/or control access to such a table, in accordance with the organization's objectives and re-identification risk assessment.

9.4.3.3 Pseudonyms derived from identifying attributes using cryptography

9.4.3.3.1 Overview of cryptography use for pseudonymization

Pseudonyms can be cryptographically derived from the values of the attributes that they replace through encryption or hashing. Such a process is sometimes referred to as "key coding" the attributes in the dataset. It is important to note that, given the appropriate key, encrypted attributes can be decrypted using the corresponding algorithm, while hashing is a one-way mathematical process.

NOTE 1 Data encryption techniques are standardised in ISO/IEC 18033 (all parts)[24]. Cryptographic hash functions are specified in ISO/IEC 10118 (all parts)[22].

Variations and combinations of these cryptographic methods can be used to better safeguard the origin of pseudonyms.

NOTE 2 For further discussion of these issues see the Article 29 working party opinion on anonymization[1].

9.4.3.3.2 Encryption

The use of encryption methods to create pseudonyms can be complex. However, these methods are effective because decryption is, in practice, infeasible in any reasonable period of time without knowledge of the appropriate cryptographic keys.

When using encryption, special measures need to be taken to safeguard the cryptographic keys from unwarranted access. These include keeping the keys separated from the data, not sharing the keys with third parties, or can involve securely erasing the keys altogether in order to prevent re-identification.

NOTE Key management principles and techniques are specified in ISO/IEC 11770 (all parts)[23].

9.4.3.3.3 Hashing

The use of a cryptographic hash-function is suitable for the purpose of pseudonymization because of their one-way and collision-resistance properties. However, hashes can be inverted if the hash algorithm is known, if the hash key is known (if a keyed hash is used – see below), and if it is possible to iterate through all possible values that can have been hashed. For example, if an 8-decimal-digit identifying attribute is hashed, it is possible to pre-compute a dictionary of the 100 million hashes of all possible 8-decimal-digit numbers. This dictionary can then be used to invert any hashed attribute.

The following additional example illustrates how small pieces of information can have a major effect. Suppose an attacker knows the unique identifier and the hash algorithm that was used. The attacker now applies the same hash algorithm to the unique identifier and looks for the result among the hashes in the database. This reveals whether or not a data principal is within the dataset/sample, and also the values of all attributes that are stored in clear for this data principal.

Using keyed hash functions (i.e. cryptographic hash functions in combination with a secret key, such as the constructions specified in ISO/IEC 9797-2[21]) involves the addition of another randomizing input which increases their suitability for pseudonymization, as they offer increased protection against unauthorized re-identification using brute force searching techniques.

Even assuming that appropriately secure hashing techniques are being used, negligence related to the use or implementation of hashing algorithms, or the sharing of keys with unauthorized parties, can result in re-identification of data.

9.5 Anatomization

The term "anatomization" refers to a category of de-identification techniques that disassociate identifiers from the remainder of the data by splitting a dataset into two tables: a table containing the existing identifiers (such as quasi-identifiers) and a table containing the remainder of the useful attributes. The attributes table is created in such a way that its rows represent the equivalence classes of records in the original table. A new attribute specifying an equivalence class is added to both tables. This attribute allows the identifiers to be mapped to their attributes (which are shared by all members of the equivalence class).

The two tables are subject to different access rights. For example, the identifiers table can be securely guarded, while the table with useful attributes can be made public.

Attribute linkage attacks are distorted by increasing the number attributes in an equivalence class. Unlike generalization or suppression, anatomization neither modifies the identifiers nor the useful attributes.

9.6 Generalization techniques

9.6.1 General

The term "generalization" refers to a category of de-identification techniques that reduce the granularity of information contained in a selected attribute or in a set of related attributes in a dataset.

Generalization techniques preserve data truthfulness at the record level. As a result, data that is de-identified using generalization is useful for cases involving traceable data principal specific patterns, such as for fraud detection, healthcare outcome assessments, etc.

Typically, the objective is to reduce the number of unique values of an attribute (or, more generally, unique combinations of values for subsets of attributes), so that each generalized value is (or subsets of values are) shared by multiple records in the resultant dataset, making it difficult to single out a data principal. Consequently, the attributes selected for generalization are typically identifying attributes, although any attribute (especially sensitive attributes) can be considered for generalization depending on the specific use case.

The output of generalization is microdata.

Generalization techniques are applicable to both numerical and non-numerical data attributes. For example: replacing "221B Baker Street London" with "London" is a generalization of a non-numerical attribute. Specific approaches to generalization of numerical data attributes are described below.

9.6.2 Rounding

Rounding involves deciding on a rounding base for a selected attribute and then rounding each value up or down to the nearest multiple of the rounding base. Whether to round up or down is decided probabilistically based on how close the observation is to the nearest multiple of a rounding base. For example, if the rounding base is 10 and 7 was observed, 7 is rounded up to 10 with probability 0,7 and rounded down to 0 with probability 0,3. Controlled rounding is also a possibility, which ensures that the sum of the rounded values is the same as the rounded value of the sum of the original data.

9.6.3 Top and bottom coding

This generalization technique sets a threshold on the largest (or smallest) value that a given attribute can take. Values that are above (or below) the threshold are replaced with a single value indicating the top (or bottom) category. This technique is applicable to attributes that are either continuous or categorical ordinal. For example, if an individual has an extremely large salary, rather than recording or reporting the exact amount the value is set to “over 100 000 USD”.

9.6.4 Combining a set of attributes into a single attribute

The granularity of information contained in a set of selected (related) attributes can be reduced by replacing them with a single attribute, whose value is computed by applying a specific function to the values of the selected attributes in the same record. The output of the aggregation of related attributes is microdata,

9.6.5 Local generalization

Local generalization involves generalizing specific values of attributes from selected records; such a procedure is used if it is possible that the attribute values, in combination with other identifying attributes, can be used to identify the data principal. Typically, local generalization is applied to remove rare values (or rare combinations of values) of indirect identifiers without modifying the remaining values of this attribute across all records. Local generalization is typically applied to numerical values with the common goal of increasing the number of records sharing the values of their identifying attributes.

9.7 Randomization techniques

9.7.1 General

The term “randomization” refers to a category of de-identification techniques in which values of an attribute are modified so that their new values differ from their true values in a random way. Such a process reduces the ability of an attacker to deduce the value of an attribute from the values of other attributes in the same data record, thereby reducing the effectiveness of inference attempts.

NOTE Another term sometimes used to describe “randomization” is “perturbation”.

Randomization techniques do not preserve data truthfulness at the record level. To achieve the chosen objectives, an effective randomization process resulting in useful data needs to be tailored on a case by case basis. Such tailoring involves both a detailed understanding of the nature of the data as well as the choice of appropriate parameters for the selected randomization techniques (and typically involves performing a statistical evaluation).

The output of randomization is microdata.

Certain randomization techniques, such as permutation, are applicable to both numerical and non-numerical data attributes. Specific approaches to randomization are described below.

9.7.2 Noise addition

Noise addition is a randomization technique that modifies a dataset by adding random values, “random noise”, to the values of a selected attribute with continuous values, while as much as possible retaining the original statistical properties of the attribute across all records in the dataset. Such statistical properties include the distribution, mean, variance, standard deviation, covariance, and correlation of the attribute.

Noise addition to a selected continuous attribute is performed by adding, or multiplying by, a stochastic or randomized number. Many different noise addition algorithms have been developed with the goal of preserving the statistical properties of the de-identified data and its usefulness for different use cases. A general review of these techniques is outside the scope of this document.

For example, the stochastic value can be chosen from a normal distribution with zero mean and a small standard deviation, such as changing a student’s numeric grade from 3,33 to 3,53.

9.7.3 Permutation

Permutation is a technique for reordering the values of a selected attribute across the records in a dataset without values’ modification. As a result, permutation retains the *exact* statistical distribution of the selected attribute across all records in the dataset.

NOTE Other terms used to describe the process of permutation are “data confusion”, “shuffling”, and “attribute substitution”.

Permutation techniques are applicable to both numeric and non-numeric values. Special considerations need to be taken to ensure that the resulting dataset appears to be consistent and realistic, because observable inconsistencies can help to reconstruct the permutation algorithm. For example, it is to be expected that men are taller than women on average; it is also to be expected that first or given names typically correspond to the listed gender.

Permutation approaches or algorithms differ both in their approach and their complexity. Some algorithms are based on repeatedly swapping values between records until all values are replaced for the selected attribute; other algorithms follow logic designed for the specific application needs. In order to preserve the correlation among the distributions of selected attributes (i.e. selected columns in a table), the same perturbation algorithm needs to be applied to all these attributes.

Knowledge of a deterministic permutation algorithm typically allows the data to be restored to its original state by back-tracking the algorithm, which makes a controlled re-identification possible. On the other hand, using a non-deterministic permutation algorithm (i.e. an algorithm that employs a degree of randomness as part of its logic) makes the process of re-identification less trivial and more resilient to re-identification attacks.

Because of this, specific organizational objectives for controlled re-identification as well as appropriate technical and organizational measures to safeguard the knowledge of the algorithms from unwarranted access need to be taken into consideration while choosing or designing the algorithm.

9.7.4 Microaggregation

The term “microaggregation” refers to a category of de-identification techniques that replace all values of continuous attributes with their averages computed in a certain algorithmic way. For each continuous attribute (or for a selected set of continuous attributes), all records in the dataset are grouped such that the records with closest values of the attribute (or attributes) belong to the same group and there are at least k records in each group, for a sufficiently large value of k. The new value of each attribute is then computed to be the average of the attribute’s values in the group. The closer the values in each group are, the more data usefulness is preserved.

The output of microaggregation is microdata. Microaggregation does not preserve data truthfulness.

Microaggregation techniques vary in the way that the attributes are selected, the way that the closeness between the values and across attributes is calculated, and other considerations – such considerations are outside the scope of this document.

9.8 Synthetic data

Synthetic data is an approach to generating microdata artificially to represent a predefined statistical data model. By definition, a synthetic dataset does not contain any data collected from or about existing data principals, but looks realistic for the intended purposes. Synthetic data fitting the original data too closely can reveal information about genuine data principals, such as their personal data.

There are various ways to create synthetic data. Theoretically, data can be randomly generated based on a number of selected statistical properties. Key characteristics of such a model are the distributions of each attribute (overall and in subpopulations) and the internal relationships among the attributes. In practice, the generation of synthetic data can involve multiple or continuous transformations on real datasets using randomization techniques (9.7) and sampling (9.1.2), as described in this document. Typically, synthetic data is used for testing tools and applications.

Synthetic data can be used for developing queries. In some applications synthetic data can be used as a surrogate for real data: in these cases, the data curator should reproduce queries performed on synthetic data on actual data, to ensure that inferences drawn on the synthetic data are correct when drawn on real data.

The privacy guarantees of synthetic data can be evaluated using the differential privacy model.

10 Formal privacy measurement models

10.1 General

A formal privacy measurement model is an approach to the application of data de-identification techniques that enables the calculation of re-identification risk and, in some cases, provides mathematical guarantees against re-identification risk. A formal privacy measurement model reflects the context of the use case.

All the models described in Clause 10 depend on the selection of parameters based on an empirical assessment of the use case, which includes both the re-identification risks of the specific system, and the characteristics of the specific dataset. System re-identification risks depend on security and other technical and operational measures that are in place. The characteristics of a specific dataset include the likelihood of an attempted re-identification attack on the dataset (which is driven by an appetite for potential incentives) and the perceived impact on the data principals and/or the organization if the attack is successful.

Many additional models have been discussed in the academic literature; examples of some of the more prominent models are outlined in Annex E.

10.2 *K*-anonymity model

10.2.1 General

K-anonymity is a formal privacy measurement model that ensures that for each identifier there is a corresponding equivalence class containing at least *K* records. While the resulting dataset has limited (i.e. 1/*K*) linkability, it does not contain measures designed to prevent potential inference attempts.

NOTE For further information, see, for example, [52] and [54].

Some of the de-identification techniques described in this document can be used either independently or in combination with each other to satisfy the *K*-anonymity model. Suppression techniques, generalization techniques, and microaggregation can be applied to different types of attributes in a dataset to achieve the desired results.

A number of enhancements to the concept of *K*-anonymity exist, including *L*-diversity and *T*-closeness.

10.2.2 *L*-diversity

L-diversity is an enhancement to *K*-anonymity for datasets with poor attribute variability. It is designed to protect against deterministic inference attempts by ensuring that each equivalence class has at least *L* well-represented values for each sensitive attribute. *L*-diversity is not a single model but a group of models (E.7). Each model has diversity defined slightly differently, e.g. by counting distinct values or by entropy.

L-diversity can be difficult to achieve and can cause significant loss in data utility due to the implicit assumption of how values for each sensitive attribute are distributed. Its ability to protect against inferences is also limited when data values are unevenly distributed. This variant of *K*-anonymity is subject to attacks, which have led to the development of *M*-invariance and *T*-closeness.

NOTE For further information, see, for example, [40] and [42].

10.2.3 *T*-closeness

T-closeness is an enhancement to *L*-diversity for datasets with attributes that are unevenly distributed, belong to a small range of values, or are categorical. It is designed to protect against statistical inference attempts, as it ensures that the distance between the distribution of a sensitive attribute in any equivalence class and the distribution of the attribute in the overall dataset is less than a threshold *T*. This technique is useful when it is important for the resulting dataset to remain as close as possible to the original one.

Enforcing *T*-closeness can cause significant loss in data utility as it eliminates the correlations between quasi-identifiers and sensitive attributes. This variant of *K*-anonymity is not resilient to linking via composition attacks, [16].

Although not equivalent, *T*-closeness and ε-differential privacy (10.3) are related to one another (see, for example, [7]).

NOTE For further information, see, for example, [39].

10.3 Differential privacy model

10.3.1 General

Differential privacy is a formal privacy measurement model that, if incorporated in the design of a particular statistical analysis, provides mathematical guarantees that the probability distribution of the output of this analysis differs by a factor no greater than a specified parameter regardless of whether any particular data principal is included in the input dataset. The specified parameter can be used to measure the “privacy loss” that the analysis incurs every time it provides an output (10.3.4.4).

The term “privacy loss” is conventionally used in the differential privacy discipline. It does not refer to actual loss of privacy but instead to a reduction in the probability that privacy is maintained (10.3.4.4). Privacy loss refers to the cumulative knowledge that a theoretical attacker acquires over time from the output of a particular statistical analysis. Differential privacy bounds the probability that the presence or absence of any particular data principal in the dataset is able to be inferred from the de-identified dataset or from system responses. This probability bound is maintained even if the attacker has access to other, related, datasets, as long as the privacy loss is limited to a certain level.

More formally, differential privacy provides:

— a mathematical definition of privacy which posits that, for the outcome of any statistical analysis to be considered privacy-preserving, the analysis results have essentially the same statistical distribution independent of whether any given data principal is added to or removed from the dataset; and

- a measure of privacy that enables monitoring of cumulative privacy loss and setting of a "budget" for loss limit.

These two elements are applied in the design of differentially private algorithms that, when adequately implemented and used, provide a mathematically proven guarantee of privacy. Any algorithm that can be proven to meet the formal definition, can be called "differentially private".

The design and construction of a differentially private algorithm requires appropriate expertise in the field of probability and statistics, and of the theory of differential privacy. Following is an overview of how differential privacy achieves its objectives, and what some of the underlying challenges for an implementation can be. A comprehensive discussion, including selection of the appropriate probability distribution and parameters, as well as practical details for implementation, is outside the scope of this document.

NOTE For detailed treatment of these and other relevant topics on the subject, see [9], [45] and [15].

Differentially private algorithms are built by adding a certain amount of "random noise" that is generated from a carefully selected probability distribution, such that the desired usefulness of data is preserved. Random noise is added either to the outputs provided by the differentially private system to an analyst (server model), or at the user device to inputs from each data principal (local model).

Variants of the differential privacy model exist – some of the more prominent examples are described in Annex E.

10.3.2 Server model

Mechanisms that follow the "server model" for differential privacy typically preserve data in unmodified form in a secure database. In order to preserve privacy, responses to queries are only able to be obtained through a software component or "middleware", known as the "curator". The curator takes queries from system users, or from reporting software, and obtains the correct, noise-free answer from the database. However, before responding to the user or reporting software, the curator adds random noise whose magnitude is inversely proportional to the privacy loss implied by the query. In this case, "random noise" means a combination of basic primitives, including but not limited to the addition of appropriately generated random numbers, to obtain algorithms that deliver a certain degree of utility in exchange for some level of privacy loss. The curator is also responsible for keeping track of the cumulative privacy loss and ensuring that it does not exceed the privacy budget. Whether the curator should stop answering queries or take other measures, if and when the privacy budget is exhausted, is established by system-specific policies defined on a case-by-case basis.

An example of a server model use of differential privacy is an application that periodically (e.g. monthly, weekly, or some other period) publishes a report detailing consumption of a resource like water, or health services, at the town or ZIP code level. Computation of consumption means, standard deviations, histograms, etc., are executed in a differentially private manner in order to avoid enabling the inference of the actual consumption of a given user, by correlating the reports with other datasets known to the attacker, whether or not they are public. It is also possible to build differentially private versions of algorithms that are commonly used to build machine learning models; optimization, linear algebra methods like singular value decomposition (SVD), commonly used for least squares fitting, and *k*-nearest neighbours for non-supervised learning. Such algorithms can provide privacy protection to data principals whose information is part of a training dataset for a machine learning model.

NOTE See [43] and [17] for further details.

10.3.3 Local model

The local model is useful when the entity receiving the data is not necessarily trusted by the data principals, or if the entity receiving the data is looking to reduce risk and practice data minimization.

In this model, data belonging to a single data principal, or the results of computations on these data, are first randomized before they are transferred to, and stored on, a server.

Data is randomized by adding a random quantity, chosen from a specific probability distribution, to each individual datum or measurement taken from data belonging to the data principal, at the user device. As a result, the individual data is randomized and thus is not representative of the actual activity taking place in each device, so does not impact the privacy of data principals. Nevertheless, when the randomized data from a large number of data principals is aggregated and used for statistical analysis at the server, the results closely follow the collective behaviour of the population. Because noise is added before transmission, data reports from the principals can, in many instances, be stored at the server without additional privacy-protective measures, and the resulting database is able to be either shared or queried directly without the need for a curator.

One scenario where the local model for differential privacy can be useful, involves privacy preserving surveys, where each data principal's responses are randomized before being sent back to the server. The pollster can then analyse the results which, assuming the usual poll criteria for sampling where followed, should closely match the opinions of the population.

10.3.4 Key considerations for a Differentially Private System

10.3.4.1 Probability distribution

In the context of differential privacy, random noise takes the form of random numbers that are generated according to a selected probability distribution. Research literature suggests the use of: Gaussian (also known as "the bell curve"), Laplacian or Exponential probability distributions, with mean zero. The most appropriate choice for each case depends on the types of queries to be made and other specifics of the model in use.

NOTE For further details, see [11] and [15].

The parameter that determines whether higher or lower noise values are likely to be produced by the noise generator is the standard deviation, a measure of how tightly the most probable noise values are grouped around zero. The magnitude of the standard deviation for a given differentially private system is proportional to the quotient S / ε, where S represents the sensitivity of a given query, and ε represents the associated privacy budget.

10.3.4.2 Sensitivity

The sensitivity, S, of a given query or function describes the worst case scenario of how much the answer to that query or that function can change if a single data principal is removed from the database. Therefore, in order to "hide" the presence of a data principal whose data causes the greatest change to the answer, a proportional amount of noise needs to be added to all answers for that particular query or function.

10.3.4.3 Privacy budget, ε

The privacy budget, designated by the Greek letter ε, is a design choice. Selecting the best value for a given differentially private algorithm or query is not a straightforward process. Given that the standard deviation of the noise is proportional to S / ε, a larger ε results in a smaller standard deviation, typically "spending" more of the privacy budget when answers are provided to users, but also carries a greater privacy risk because smaller noise values are more likely to be added to the actual results. A smaller ε on the other hand increases the magnitude of the standard deviation, consequently increasing the likelihood that larger noise values are added to the actual results, thus providing greater privacy protection. At the time of this publication, ways of selecting the best ε for a specific scenario are still under research.

10.3.4.4 Cumulative privacy loss

Every query that the differentially private algorithm responds to, carries with it a privacy cost or privacy loss. In a well-designed differentially private algorithm, these losses are small enough that privacy is not breached, but their cumulative effect can eventually lead to a breach of privacy. In order to calculate the changes in a privacy budget, the notion of a cumulative loss from multiple queries is defined.

In a simple case, if n queries with a similar privacy cost, C, are posed to a differentially private algorithm, the total privacy budget spent is no greater than nC. (This is also true if the n queries asked are all the same question). Estimating the exact privacy cost of a given query in a general case, including scenarios when queries of different sensitivity are asked and the random noise component comes from different probability distributions, is not trivial and are described in [9].

It is important, however, to point out that exhausting the privacy budget of a specific system does not imply an automatic breach of privacy, but rather invalidation of the mathematical guarantee. Once the guarantee is invalidated, output from the algorithm can be used by an attacker to apply inference, linking and other types of re-identification approaches, which can result in a successful re-identification attack.

10.4 Linear sensitivity model

10.4.1 General

Linear sensitivity is a model that refers to a set of measures with specified parameters that indicate how close a data principal's contribution can be estimated from the aggregation of an attribute in microdata, regardless of whether the released data contains the aggregated value (macrodata release) or the individual values collectively conforming to the model (microdata release). Specified parameters are used to ensure a "sufficient distance" from the true value, to make it difficult for an attacker to estimate too closely the contribution of any specific data principal based on knowledge of another data principal. Typically, a combination of linear sensitivity measures are used, addressing different potential risks.

More formally, a linear sensitivity measure is the linear weighted sum of all values of a selected attribute across the records in a dataset (e.g. based on an equivalence class, or an equivalence class and a sensitive attribute), in which the values are ordered in decreasing order, and the corresponding sequence of weights are also decreasing. The linear sensitivity measure is considered "sensitive" when the linear weighted sum is strictly greater than 0.

The theory of linear sensitivity measures has driven methodology and practice for several decades. Although typically described and used in the release of macrodata, they are also used in the release of microdata. Commonly used linear sensitivity measures include the threshold rule, dominance rule, and ambiguity rule.

10.4.2 Threshold rule

Inferences can often be drawn where there is an insufficient number, n, of data principals in a selected attribute across the records in a dataset. This can be the number of data principals in an equivalence class, or the number of data principals contributing to the aggregation of a sensitive attribute. A minimum number of data principals in a selected attribute is define by a threshold, n, below which the number of data principals in the selected attribute is deemed sensitive.

NOTE 1 The threshold rule is also known as the n-rule.

NOTE 2 For microdata, a threshold n corresponds to the parameter K in K-anonymity for a selected subset of identifying attributes (10.2), and a threshold n corresponds to the value L in L-diversity (10.2.2).

10.4.3 Dominance rule

Aggregation can be "dominated" by a small number of the largest values in a selected attribute across the records in a dataset, and therefore be subject to inference. A specific number, *n*, of the largest attribute values can be compared against a maximum percentage, *k*, of the aggregation. The aggregation is deemed sensitive if the *n* largest attribute values contribute more than the maximum percentage, *k*, of the aggregation. This model does not, however, provide consistent protection for the aggregation of all attributes being released.

NOTE The dominance rule is also known as the (*n*,*k*)-rule.

10.4.4 Ambiguity rule

The contribution of one value by a data principal in a selected attribute in aggregation can be estimated before and after data has been released. The information gained through data release can be deemed too revealing because of the inferences that can be drawn. Estimating a data principal's contribution to aggregation to within a certain percentage, *p*, before data release can be compared to estimating a data principal's contribution to aggregation to within a certain percentage, *q*, after data release. The aggregation is deemed sensitive if the information gain (p / q) before and after data release is greater than a maximum information gain. This model provides consistent protection across all aggregation, but requires quantification of prior knowledge.

NOTE The ambiguity rule is also known as the *p*/*q*-rule.

11 General principles for application of de-identification techniques

11.1 General

This document describes a broad range of de-identification techniques for reducing the risk of re-identification. The process of selecting the de-identification techniques (which are characterized in Clause 9) needs to be tailored to each specific use case. Although there is no best or standard way in which the selection process can be done in all cases, the factors are presented in a logical order that can be used in practice as a part of a data processing system design. Not all the listed steps are relevant in every use case.

NOTE Existing sector-specific guidelines specify detailed processes for the selection of de-identification techniques. For example, ISO 25237[27], addresses the health sector. Specification of such processes is outside the scope of this document.

In a specific data processing system, feasible technical and organizational measures are implemented to balance the need for preservation of data usefulness with the need for data de-identification (see Clause 12). They provide an essential context for the choice of the degree to which the data is de-identified. In turn, this degree of de-identification can be achieved by tuning the selection of identifying attributes and the techniques as a part of the selection process described in Clause 11.

Performing data minimization, i.e. limiting the data to what is directly relevant and necessary to accomplish a specified purpose, at the earliest possible stage typically makes the task of data de-identification easier.

In some cases, quantifiable guarantees against the risk of re-identification need to be achieved. This can be done by implementing one of the formal privacy measurement models described in Clause 10. Suggested designs and implementations of such models tailored to different use cases and objectives are described in the existing literature. They are not presented in this document.

11.2 Sampling considerations

Performing a random sampling of individual records in the dataset as the first step in the de-identification process adds uncertainty about the presence of a specific data principal in the dataset and thus can increase the effectiveness of subsequent de-identification techniques. Sampling can also simplify the calculation of aggregated values of very large data populations.

11.3 Aggregated vs. microdata

The output of de-identification techniques can be of two main types: *aggregated data*, which is data representing a group of data principals, such as a set of statistical properties of the group or *microdata*, which is a dataset comprised of records about individual data principals.

Since microdata contains records about individual data principals there is always some risk of successful re-identification. In this sense, data aggregation is the more robust de-identification technique, but it is only applicable to certain types of data and is useful in a limited set of use cases. Generation of aggregated data is described in 9.1.3. Therefore, it is beneficial to decide on whether data aggregation is feasible and sufficient for the use case in hand. If the answer is positive, the selection process can stop.

11.4 Classification of attributes

All attributes need to be classified according to the technical model defined in Clause 6. The attributes in a particular dataset that can be classified as identifying attributes (and in particular those that are direct identifiers) first need to be determined. In certain contexts, classification into indirect identifiers can depend on the assessment of data available outside the dataset. Classification into sensitive attributes depends on the meaning and the nature and type of data.

All known re-identification methods can be considered for the selection of a subset of attributes to be subjected to de-identification techniques in a given operational context (e.g. all direct identifiers only, all unique identifiers and most indirect identifiers, or all identifying attributes and all sensitive attributes). Collectively, the decisions made in the selection process determine the extent to which the de-identified data can be associated with individual data principals.

Certain conditions can help inform the classification of attributes: the stability of attribute values over time so that the values occur consistently in relation to the data principal; the variability of an attribute to distinguish among data principals in a dataset; and the attacker's knowledge in a given operational context.

11.5 Handling of direct identifiers

The two common ways of dealing with direct identifiers is either removing them from the dataset (using masking as described in 9.3.2) or replacing their values with pseudonyms (using pseudonymization as described in 9.4). Unlike masking, pseudonymization preserves the ability of linking records belonging to the same data principal even after the data is de-identified.

11.6 Handling of remaining attributes

In order to protect data and to avoid re-identification attacks using linking and inference approaches, the unique combinations of indirect identifiers and sensitive attributes need to be removed from the dataset.

This can be achieved by selectively applying generalization (9.6), randomization (9.7), suppression (9.3) or encryption-based (9.2) de-identification techniques to the remaining attributes.

The outliers (i.e. values of attributes that are much larger or smaller than most of the other values of this attribute in the dataset) can be removed at either local (9.3.3) or record (9.3.4) level. For statistical or aggregation purposes, record suppression typically results in a greater distortion of data than local suppression.

To preserve data usefulness for certain use cases (such as fraud detection or healthcare outcome assessments), de-identified data needs to remain truthful. Data is considered truthful if it is factual and has not been accidentally or deliberately distorted. Truthful data cannot contain incorrect or artificially generated information. Generalization typically preserves data truthfulness. Suppression preserves data truthfulness at the record level but not at the aggregated level, because the statistical analysis is distorted by the removal of certain records or attribute values. Conversely, randomization does not preserve data truthfulness at the record level, but does preserve it at the aggregated level because statistical functions computed per attribute give the same result as when no randomization is applied.

Both generalization and randomization techniques use mathematical or logical operations on attribute values. Since only particular operations can be performed on particular value types, attributes need to be classified into numerical, categorical, or ordinal. Numerical values can be discrete or continuous. Discrete values are countable. Continuous values typically represent measurements in terms of real numbers and are not countable. Categorical values represent characteristics and do not have a mathematical meaning. Ordinal values also represent characteristics; however, unlike categorical data, the categories (typically represented by numbers) can be ordered.

NOTE Another term sometimes used to describe “categorical” is “qualitative”.

Once attributes are classified according to their value type, only de-identification techniques suitable for that value type can be considered for each of the attributes. Subsequently, if multiple suitable techniques are found to be available, the usefulness of the resulting data is a major factor in the selection of the best fitting de-identification technique.

11.7 Privacy guarantee models

Typically, a de-identification technique alone cannot provide quantifiable guarantees against re-identification attacks. To achieve such a goal, a range of “formal privacy measurement models” have been designed, each relying on a different selection of one or more de-identification techniques and each providing its own approach to calculating the risk of re-identification in mathematical terms. These models are described in Clause 10.

12 Additional technical or organizational measures

12.1 General

In order to balance the need for preservation of data usefulness with the need for data de-identification, de-identification techniques and models typically need to be accompanied by technical and other organizational and operational measures to enhance their effectiveness. These measures include authentication, authorization, encryption, legal agreements, regulations to protect the data at rest and in transit, and appropriate training, including a code of conduct, for personnel having permanent or frequent access to personal data. Their choice and design is dependent on the environment in which data is collected, processed, or shared. Encryption keys, seed for pseudorandom number generation, mapping tables, dictionaries, configurations, and other information used as part of the de-identification process should also be appropriately handled, stored and retained/deleted.

Factors playing an essential role in the design of technical and other organizational measures are: the data flow, and the means used to access the data. These factors are orthogonal to the characteristics of the de-identified data. Therefore, if more than one de-identification technique or model is deemed to be applicable to a specific use case (for example, as described in Clause 11), the feasibility of the deployment of technical and other organizational measures to protect de-identified data for each selected technique or model need to be assessed before deciding on the most practical approach to de-identification for a given use case.

Detailed discussion of the implementation of technical and other organizational measures is outside the scope of this document.

12.2 Data flow scenarios

Typically, an organization sets its objectives for privacy and security measures, including data de-identification, in accordance with its business needs and relevant laws and regulations, and in support of widely held privacy principles.

As a result, organizations are incentivized to implement data de-identification techniques in various data flow scenarios including:

- collection of data about a large number of individuals, such as for research purposes;
- collection of data under a contract, such as for processing users' data by a PII processor;
- collection of data automatically generated by sensors and that is directly or indirectly related to a natural person, device, etc.;
- sharing of data within the organization, such as using data by multiple applications;
- sharing of data with a third party, such as for secondary use by another PII controller;
- public sharing (or release) of data;
- testing/training of models or systems using personal data.

The nature of the relationships between the parties participating in the data flow affects whether data de-identification needs to be performed before its collection, after its collection but before its storage, or only before it is shared with the next party in the data flow. This decision, in turn, affects the feasibility of technical and other organizational measures to enhance the effectiveness of a particular de-identification technique in each use case.

12.3 Access to de-identified data

The implementation of de-identification techniques and models can vary depending on the technical means provided to access the resulting data. The distinction is in whether the de-identified dataset can be published or retrieved as a whole, or the information can only be retrieved through responses to predefined data queries.

NOTE In the latter case, in addition to security measures, the responses can include data manipulation specific to the de-identification techniques or the model.

Although all de-identification techniques and models described in this document can be designed to support either approach to access the data, the required technical and other organizational measures can differ significantly between the two cases.

12.4 Controlled re-identification

Controlled re-identification is a process that can be built into the data processing system to ensure that deterministic data re-identification is possible using specified technical and organizational measures. An example of the need for controlled re-identification is to support data principal requests for access to their own data. Further discussion regarding possible circumstances in which controlled re-identification is performed, and the techniques by which it is achieved, are outside the scope of this document. The notion of controlled re-identification is introduced solely to clarify how the terminology defined and used in this document relates to certain terms used in the prior art (see Annex B).

Annex A
(informative)

Summary of de-identification tools and techniques

Table A.1 summarizes the key properties of the de-identification techniques and tools described in this document, and identifies risks that can be reduced through the use of each technique.

Table A.1 — Properties of de-identification tools, techniques and models

Technique name	Data truthfulness at record level	Applicable to types of values	Applicable to types of attributes	Reduces the risk of		
				Singling out	Linking	Inference
Statistical tools						
Sampling						
Aggregation	N/A	Continuous, discrete	All attributes	Yes	Yes	Yes
Cryptographic tools	Yes					
Deterministic encryption	Yes	All	All attributes	No	Partially	No
Order-preserving encryption	Yes	All	All attributes	No	Partially	No
Homomorphic encryption	Yes	All	All attributes	No	No	No
Homomorphic secret sharing	Yes	All	All attributes	No	No	No
Suppression	Yes					
Masking	Yes	Categorical	Local identifiers	Yes	Partially	No
Local suppression	Yes	Categorical	Identifying attributes	Partially	Partially	Partially
Record suppression	Yes	N/A	N/A	Partially	Partially	Partially
Sampling	Yes	N/A	N/A	Partially[a]	Partially	Partially
Pseudonymization	Yes	Categorical	Direct identifiers	No	Partially	No
Generalization	Yes	All, subject to meaning	Identifying attributes			
Rounding	Yes	Continuous	Identifying attributes	No	Partially	Partially
Top/bottom coding	Yes	Continuous, ordinal	Identifying attributes	No	Partially	Partially
Randomization	No		Identifying attributes			
Noise addition	No	Continuous	Identifying attributes	Partially	Partially	Partially

[a] If the data principal record is not included in the sample.

[b] Unless *K*-anonymity is implemented using microaggregation.

N/A Not applicable.

Table A.1 *(continued)*

Technique name	Data truthfulness at record level	Applicable to types of values	Applicable to types of attributes	Reduces the risk of		
				Singling out	Linking	Inference
Permutation	No	All	Identifying attributes	Partially	Partially	Partially
Micro aggregation	No	Continuous	Indirect identifiers, and all attributes	No	Partially	Partially
Differential privacy	No	All	Identifying attributes	Yes	Yes	Partially
K-anonymity	Yes[b]	All	Quasi identifiers	Yes	Partially	No

[a] If the data principal record is not included in the sample.

[b] Unless *K*-anonymity is implemented using microaggregation.

N/A Not applicable.

Annex B
(informative)

Prior art terminology

B.1 General

The purpose of this annex is to clarify how the terminology used in this document relates to terms being used in prior art.

This document reuses prior terminology wherever it has been defined and used consistently. However, in some cases the definitions provided in this document are not always identical to earlier definitions because the prior art sometimes uses terminology inconsistently and not in accordance with the state of the art; the terms used in this document have been formulated to capture the latest understanding in the field of data de-identification, and avoids using terms that have been used inconsistently in prior art as discussed below.

The prior art tends to entangle technical de-identification techniques with technical and other organizational measures implemented to enhance the effectiveness of de-identification. The terminology defined in this document has been designed to allow the discussion of de-identification to be distinguished from technical and other organizational measures, and as a result permits this document to focus on de-identification techniques.

B.2 Definitions of individual terms

This document does not use the term “reversible” or its derivatives (i.e. irreversible, reversibility, etc.) because this term has been used to denote organizational measures being taken to allow or disallow controlled re-identification (as shown in Table B.1 below) as well as to describe mathematical properties of transformation functions (e.g. of cryptographic hash-functions).

This document also does not use the term “anonymize” or its derivatives (i.e. anonymisation, anonymization, anonymised, anonymized, anonymous, anonymity, etc.) because the term has been used in the past to convey a range of different meanings, as shown in Table B.1 below.

Table B.1 — Mapping of de-identification terminology to the prior art

Term used in this document	ISO 25237[27]	ISO 29100[29]	ICO 2012[20],	Article 29 2014[1],
De-identification	De-identification, Anonymization	Anonymization	Anonymisation	N/A
Masking	N/A	N/A	Anonymisation	N/A
Pseudonymization with controlled re-identification	Pseudonymization reversible	Pseudonymization	Anonymisation	Pseudonymisation
Pseudonymization without controlled re-identification	Pseudonymization irreversible	Anonymization	Anonymisation	Pseudonymisation
Randomization	N/A	N/A	Anonymisation	Anonymisation
Generalization	N/A	N/A	Anonymisation	Anonymisation
Differential Privacy	N/A	N/A	N/A	Anonymisation

N/A Not applicable.

B.3 Relationship to ISO/IEC 19944

Table B.2 shows how the privacy enhancing de-identification techniques described in this document, when used to process data containing PII, can result in a state of data characterized by one or more of the data identification qualifiers listed in ISO/IEC 19944[26].

NOTE Use of the techniques listed in this table are not the only way of achieving the data identification states described by the qualifiers.

Table B.2 — Mapping between ISO/IEC 19944 and ISO/IEC 20889 terminology

ISO/IEC 19944 data identification qualifiers describing state of data	ISO/IEC 20889 Privacy enhancing data de-identification techniques whose application leads to the corresponding state of data
Identified data	original, unprocessed data containing identifiers; in other words, no de-identification techniques are applied yet; for the other qualifiers, the identifiers are removed (masked)
Pseudonymized data	data processed using pseudonymization techniques with controlled re-identification possible/implemented
Unlinked pseudonymized data	data processed using pseudonymization techniques with no controlled re-identification allowed
Anonymized data	data processed using generalization and/or randomization techniques
Aggregated data	data processed using aggregation techniques

B.4 Relationship to Statistical Disclosure Control terminology

Statistical Disclosure Control (SDC) [also known as Statistical Disclosure Limitation (SDL)] is a discipline that was originally developed for controlling the disclosure risk when publishing census and survey data. Subsequently, SDC has been applied to other, similar, use cases, such as the publication of health records and medical statistics. As such, SDC significantly contributed to the broader practice of data de-identification, which is the focus of this document. Many SDC techniques are, in fact, de-identification techniques and are included in this document.

Since the establishment of SDC, both refinements to known de-identification techniques and new de-identification models have been developed to address additional use cases for data de-identification. As a result, new terms, rooted in additional disciplines, have been introduced to the field of data de-identification. This document normalizes the resulting general-purpose terminology, which has become the predominantly used terminology across a range of industries and academia.

In some cases, the terminology used in this document overlaps with SDC terminology. Table B.3 below lists such cases where a simple mapping is possible.

Table B.3 — Mapping to SDC terminology

Terminology in this document	SDC terminology
Attribute	Variable
Attacker	(Data) Intruder
Data usefulness	Data utility
Generalization	Global recoding
Indirect identifier	Key variable

In addition, in some cases similar terminology is used to refer to different concepts as described below.

In SDC, the term “tabular data” (also known as “aggregate tabular data” or “aggregate data”) refers to the practice of creating (and subsequently publishing) tables solely containing statistics about the original values of the attributes, such as frequency count table. In this document, the terms “aggregation” and “aggregated data” refer to statistical computations on the original data, neither implying the format nor specifying the means to access the resultant computations.

The term “tabular” is not used in this document to avoid confusion between its specific meaning in the field of statistical disclosure control and its English meaning of “in a structured form”.

Annex C
(informative)

De-identification of free-form text

C.1 General

The purpose of this annex is to describe a number of existing approaches to de-identification of free-form text and clarify how selected de-identification techniques described in this document are applicable for de-identification of free-form text.

In addition to the complexity involved with de-identification of structured data, the nature of free-form text creates additional challenges specific to this format.

Firstly, in free-form text, the semantics of each word is not specified as a part of the dataset. Therefore, segmentation of text by parsing is typically performed before any kind of de-identification can take place.

Secondly, multiple data principals can be described or referred to in a given text without obvious association between each principal and its characteristics. Consequently, the de-identification models presented in this document, which enable the calculation of re-identification risk, are typically not applicable to de-identification of free-form text.

Where a de-identification approach involves the removal of information from a free-form text document, the procedures and technology described in ISO/IEC 27038[29] can be used as appropriate.

The following clauses describe several known de-identification techniques that are specific to free-form text.

NOTE In the existing literature, the term "data segmentation" is sometimes used to describe the whole process containing any of the de-identification techniques described below. The text in this annex distinguishes between multiple stages and different techniques in a process.

C.2 Annotating

Annotating (or named entity recognition) is not a de-identification technique on its own. Rather, it is a process that needs to be performed on free-form text before any type of de-identification can take place.

The objective of annotating is to divide logically a given free-form text into segments and annotate the segments with their semantics whenever possible. A segment can be a single word or a string of words. A segment can hierarchically contain other segments.

Annotating can be either rule-based or dictionary-based. Dictionary-based annotating can be enhanced by using learning mechanisms.

A basic example of rule-based annotating is dividing a text into tokens (e.g. words) in conformance to the rules of syntactic grammar. For example, it can be done by defining whitespace characters as the delimiters between the tokens. Another example of rule-based annotating is dividing the text into a defined number of words (also known as *n*-gram text).

More advanced rules can be defined by using regular expressions. For the purpose of de-identification, each regular expression represents a structure of a commonly used attribute, such as a date or an e-mail address.

Dictionary-based annotating is based on matching the text to a set of known segments. For the purpose of de-identification, the dictionary contains the values of known identifying, sensitive and other attributes. Additionally, the dictionary can contain text segments serving the purpose of the application using the de-identified data. The dictionary can be populated upfront and can be updated during the process of annotating.

The process of annotating can be performed in different ways. That is, the text can be matched against the available patterns (i.e. regular expressions and segments) sequentially, in parallel, or as a combination of both. The available patterns can be prioritized according to different considerations. To achieve dynamic prioritization, learning mechanisms can be used during the annotating process.

The output of annotating is the original text with each delimitated segment being marked with its semantics whenever possible.

Use of machine learning, natural language processing (NLP), and other forms of context analysis can be used to improve the understanding and quality of classifying sensitive data. It is also important for context-specific models such as language-specific, healthcare, financial, etc.

C.3 Conversion of data to the form of a table

In some cases, as a result of parsing, it is possible to arrange the delimitated segments as a table with columns and rows according to their semantics and discard the rest of the data. In such case, the parsed file has effectively been converted into a structured dataset that is able to be de-identified using any of the techniques presented in this document.

In practice, due to complexity of a given free-form text, often it is impossible to convert the text into a structured dataset without losing much of its value. This limits the selection of the de-identification techniques that are applicable to free-form data as described in C.4 below.

C.4 Scrubbing

The objective of scrubbing is to produce a free-form text that preserves the structure and much of the content of the original free-form text, but which neither contains the identifiers of data principals nor it contains any presumably sensitive information about them.

Scrubbing effectiveness relies on the ability of the parsing process to assign semantics to as many recognisable segments as possible. Scrubbing is performed by applying selective de-identification techniques described in this document to the recognized attributes according to their semantics.

Multiple data principals can be described or referred to in a given text without obvious association between a principal and its attributes. Performing de-identification on each recognized attribute independently from others limits the type of de-identification techniques applicable to free-form text. As a result, masking (9.3.2) and pseudonymization (9.4) are the two de-identification techniques being mostly used for scrubbing. The specific choice of the de-identification technique to be applied to each recognized attribute depend on its semantics and type.

De-identification can be performed on-the-fly during the parsing. Alternatively, during the parsing the recognized segments can be marked with their meaning (i.e. semantics) and de-identified after the parsing is completed.

C.5 Segmentation

The objective of free-form text segmentation is to produce (and potentially share or publish) a set of text segments such that the segments:

- neither contain the identifiers of data principals nor any presumably sensitive information about them;
- cannot be put together to reconstruct the original text.

Parsing and scrubbing together are used to produce a useful set of segments. The length of the output segments can be one of the parameters to the parsing process. The shorter the segments are, the less is the risk of re-identification.

Risk of re-identification can be reduced by only releasing the segments that are also found in publicly available text content, such as large text corpuses from Web sites.

C.6 Aggregation

This approach is similar to the aggregation de-identification techniques presented in this document. The objective of free-form text aggregation is to produce a set of text segments (as described in C.5) with statistics about them.

The frequency of data segments in a given text is an example of straightforward statistics. Statistics related to sensitive attributes removed (or otherwise de-identified) during de-identification of the segment are another example of aggregation functions.

Annex D
(informative)

Normalization of structured data

If, in practice, multiple rows exist corresponding to a single data principal, then it is possible to merge them. However, not all the formal privacy measurement models described in this document work as expected, or in an optimal way, depending on the way the data is merged. Some examples of how data can be merged are presented.

In a binary view of the data, attributes that have multiple values across records for a single data principal are transformed into attributes themselves. For example, the attribute “month of visit” is transformed into attributes representing the months of the year (i.e. “January”, “February”, ..., “December”). Entries in the binary view are then just indicators of whether the data principal has this attribute or note. However, the relationship between attribute values can be lost in this representation. For example, the attribute “year of visit” is transformed into attributes representing each year, independent of the attribute “month of visit” (i.e. which “year” is associated with which “month” is lost). Alternatively, attribute values can be combined.

In a transposed view of the data, the maximum number of attribute values for a single attribute and a single data principal determines the number of new ordered attributes that are created. For example, if the attribute “month of visit” has at most three values for a single data principal, the transposed view has three attributes “month of visit 1”, “month of visit 2”, and “month of visit 3”. Entries in the table are then the multiple values across records for a single data principal’s attribute values themselves. The transposed view has fewer attributes than the binary view, preserve the order of records, and maintain the relationship between the multiple attribute values for a single data principal. However, depending on the formal privacy measurement model or the operational context, attribute values can be combined.