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/IEC 2018



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2018

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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 ..... **v**
Introduction ..... **vi**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Symbols and abbreviated terms **5**
5 Overview **6**
6 Technical model and terminology **6**
7 Re-identification **8**
7.1 General 8
7.2 Re-identification attacks 8
8 Usefulness of de-identified data **10**
9 De-identification techniques **10**
9.1 Statistical tools 10
9.1.1 General 10
9.1.2 Sampling 10
9.1.3 Aggregation 11
9.2 Cryptographic tools 11
9.2.1 General 11
9.2.2 Deterministic encryption 11
9.2.3 Order-preserving encryption 12
9.2.4 Format-preserving encryption 12
9.2.5 Homomorphic encryption 13
9.2.6 Homomorphic secret sharing 13
9.3 Suppression techniques 14
9.3.1 General 14
9.3.2 Masking 14
9.3.3 Local suppression 15
9.3.4 Record suppression 15
9.4 Pseudonymization techniques 15
9.4.1 General 15
9.4.2 Selection of attributes 15
9.4.3 Creation of pseudonyms 16
9.5 Anatomization 17
9.6 Generalization techniques 17
9.6.1 General 17
9.6.2 Rounding 18
9.6.3 Top and bottom coding 18
9.6.4 Combining a set of attributes into a single attribute 18
9.6.5 Local generalization 18
9.7 Randomization techniques 18
9.7.1 General 18
9.7.2 Noise addition 19
9.7.3 Permutation 19
9.7.4 Microaggregation 19
9.8 Synthetic data 20
10 Formal privacy measurement models **20**
10.1 General 20
10.2 ***K*-anonymity model** 20
10.2.1 General 20

© ISO/IEC 2018 – All rights reserved

10.2.2 ***L*-diversity** 21
10.2.3 ***T*-closeness** 21
10.3 Differential privacy model 21
10.3.1 General 21
10.3.2 Server model 22
10.3.3 Local model 22
10.3.4 Key considerations for a Differentially Private System 23
10.4 Linear sensitivity model 24
10.4.1 General 24
10.4.2 Threshold rule 24
10.4.3 Dominance rule 25
10.4.4 Ambiguity rule 25

11 General principles for application of de-identification techniques 25
11.1 General 25
11.2 Sampling considerations 25
11.3 Aggregated vs. microdata 26
11.4 Classification of attributes 26
11.5 Handling of direct identifiers 26
11.6 Handling of remaining attributes 26
11.7 Privacy guarantee models 27

12 Additional technical or organizational measures 27
12.1 General 27
12.2 Data flow scenarios 27
12.3 Access to de-identified data 28
12.4 Controlled re-identification 28

Annex A (informative) **Summary of de-identification tools and techniques 29**

Annex B (informative) **Prior art terminology 31**

Annex C (informative) **De-identification of free-form text 34**

Annex D (informative) **Normalization of structured data 37**

Annex E (informative) **Overview of approaches to formal privacy measurement models 38**

Bibliography 43

 © ISO/IEC 2018 – All rights reserved

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.

© ISO/IEC 2018 – All rights reserved

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.

© ISO/IEC 2018 – All rights reserved

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)

© ISO/IEC 2018 – All rights reserved

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, $\Pr\left[M\left(D_1\right) \in S\right] \leq e^{\varepsilon} \Pr\left[M\left(D_2\right) \in S\right]$, 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)

© ISO/IEC 2018 – All rights reserved

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)

© ISO/IEC 2018 – All rights reserved

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.

© ISO/IEC 2018 – All rights reserved

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

© ISO/IEC 2018 – All rights reserved