GB

NATIONAL STANDARD OF THE

PEOPLE'S REPUBLIC OF CHINA

ICS 35.030

CCS L 80

GB/T 42460-2023

Information security technology - Guide for evaluating the effectiveness of personal information de-identification

信息安全技术 个人信息去标识化效果评估指南

Issued on: March 17, 2023 **Implemented on: October 01, 2023**

Issued by: State Administration for Market Regulation;
Standardization Administration of the People's Republic of China.

Table of Contents

Foreword 3
Introduction 4
1 Scope 5
2 Normative references 5
3 Terms and definitions 5
4 Grading of personal information de-identification effectiveness 7
5 Evaluation process for effectiveness of personal information de-identification 8
6 Evaluation implementation 9
6.1 Evaluation preparation 9
6.2 Qualitative evaluation 10
6.3 Quantitative evaluation 10
6.4 Formation of evaluation conclusions 11
6.5 Communication and negotiation 11
6.6 Evaluation process documentation management 11
Annex A (informative) Examples for direct identifiers 13
Annex B (informative) Examples for quasi-identifiers 14
Annex C (informative) Identification of quasi-identifier 15
Annex D (informative) Examples for de-identification effectiveness evaluation based on K-anonymity model 17
Bibliography 25

Information security technology - Guide for evaluating the effectiveness of personal information de-identification

1 Scope

This document provides guidelines for grading and evaluating the effectiveness of personal information de-identification.

This document applies to personal information de-identification activities. It is also applicable to personal information security management, supervision and evaluation.

2 Normative references

The following referenced documents are indispensable for the application 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.

GB/T 25069-2022, *Information security techniques -- Terminology*

GB/T 35273-2020, *Information security technology -- Personal information security specification*

GB/T 37964-2019, *Information security technology -- Guide for de-identifying personal information*

3 Terms and definitions

For the purposes of this document, the terms and definitions defined in GB/T 25069-2022, GB/T 35273-2020, GB/T 37964-2019 as well as the followings apply.

3.1 personal information

Various information related to identified or identifiable natural persons recorded electronically or otherwise.

NOTE: Anonymized information is not included.

[Source: GB/T 35273-2020, 3.1, modified]

3.2 personal information subject

The natural person identified or associated with the personal information.

3.9 completely public sharing

Once the data is released, it is difficult to recall, and it is generally released directly through the Internet.

[Source: GB/T 37964-2019, 3.12]

3.10 controlled public sharing

The use of data is constrained by the data use protocol.

[Source: GB/T 37964-2019, 3.13]

3.11 enclave public sharing

Share within physical or virtual jurisdictions. Data cannot be exported outside the territory.

[Source: GB/T 37964-2019, 3.14]

3.12 re-identification risk; identifiability

The probability that the subject of personal information can be identified from the data.

3.13 equivalence class

A collection of rows in microdata where all quasi-identifier attribute values have the same value.

3.14 acceptable risk threshold

The set re-identification risk threshold value.

NOTE: When the re-identification risk is greater than this value, mitigation measures (including de-identification processing) and emergency measures need to be taken to keep the risk within a controllable range.

4 Grading of personal information de-identification effectiveness

Based on whether the data can directly identify the subject of personal information, or how likely it is to identify the subject of personal information, the identifiability of personal information is graded into four levels, see Table 1, used to distinguish the effectiveness of de-identification of personal information.

Table 1 -- 4 levels of personal information identifiability

Grading	Grading basis

c) Form an evaluation team, including personal information protection compliance experts, de-identification technical experts, and relevant business experts.

d) Carry out preliminary research, including detailed research on the data usage environment.

e) Determine the evaluation basis, including relevant laws, regulations and standards.

f) Determine the re-identification risk calculation scheme and acceptable risk threshold:

1) The re-identification risk calculation scheme considers both the dataset and the context in which it is used. It can be based on K anonymous model or differential privacy model, etc.

2) The acceptable risk threshold meets the corresponding safety requirements and meets the application needs.

g) Develop an evaluation plan.

6.2 Qualitative evaluation

Qualitative evaluation includes:

a) Identify the identifier according to 5.3 in GB/T 37964-2019. Form a list of identifiers (including direct identifiers and quasi-identifiers).

b) Determine whether the dataset contains identifiers in the identifier list. If it does not contain any identifiers, it is rated as level 4 and the evaluation ends; otherwise continue.

c) Determine whether the dataset has eliminated direct identifiers from the identifier list. If it contains the direct identifiers in the list, it is rated as level 1, and the evaluation ends; otherwise, further quantitative evaluation is carried out.

6.3 Quantitative evaluation

Quantitative evaluation includes:

a) Quantitatively calculate the re-identification risk. Carry out re-identification risk calculation according to the re-identification risk calculation scheme determined in 6.1f).

b) Compare the calculated re-identification risk results with acceptable risk thresholds. If the re-identification risk result is less than the acceptable risk threshold, it is rated as level 3; otherwise, it is rated as level 2, and the evaluation ends.

See Annex D for the re-identification risk calculation scheme and evaluation example based on the K-anonymity model.

6.4 Formation of evaluation conclusions

The formation of evaluation conclusions includes:

a) Combining the results of qualitative and quantitative evaluations, a grading conclusion for de-identification effectiveness is formed.

b) The conclusion is approved by management officials.

6.5 Communication and negotiation

During the evaluation process, maintain communication with relevant parties (including data providers, data receivers, etc.) and record the communication content, including:

a) Confirmation of understanding of data sharing purpose and data sharing environment;

b) Establishment of notification mechanism for major data environment changes;

c) Mutual exchange of information and views on re-identification risk metrics;

d) Opinions expressed by interested parties on the risk of re-identification;

e) Plan for regular/irregular reassessment.

6.6 Evaluation process documentation management

Evaluation process documentation management includes the following.

a) Evaluation process documents include the basis, reference and generated process documents and result documents during the evaluation process, including but not limited to:

1) Evaluation plan: including the data set to be evaluated, the environment for data use, evaluators, evaluation methods, formation of evaluation results and implementation progress, etc.;

2) Identifier identification report: the process and results of identifier identification;

3) Re-identification risk calculation scheme: the re-identification risk calculation scheme and the determination process and results of the acceptable threshold

Annex A

(informative)

Examples for direct identifiers

Any identification number, characteristic or code that uniquely identifies an individual in a particular context is a direct identifier. Common direct identifiers include, but are not limited to:

a) Name;

b) Citizenship number;

c) Passport number;

d) Driver's license number;

e) Detailed residential address;

f) Email address;

g) Telephone numbers (including mobile phone numbers and landline numbers);

h) Fax number;

i) Bank account;

j) Vehicle identifier and serial number (including license plate number);

k) Social security number;

l) Health card number;

m) Medical record number;

n) Device identifier and serial number;

o) Biometric identification codes (including identification codes such as fingerprints and voiceprints);

p) Full face image and any other comparable images;

q) Account number, certificate number or license number;

r) Internet Protocol (IP) addresses.

Annex C

(informative)

Identification of quasi-identifier

C.1 Considerations for identifying quasi-identifiers

Quasi-identifiers are attributes in microdata. Combining with other attributes, it can uniquely identify the subject of personal information. Usually, the information in the quasi-identifier can be known by the acquaintances of the personal information subject or exist in some kind of database.

There are usually some simpler ways of doing things to identify quasi-identifiers. For example, all other attributes except direct identifiers are regarded as quasi-identifiers. This method does not consider the possibility of attributes being combined by data receivers and other background knowledge (other external data sources) for association attacks. A plethora of quasi-identifiers may be formed. If the K-anonymity method is used for processing, a large amount of information may be lost, and the de-identified data cannot support the original application purpose. Another approach is to consider the possibility of correlation attacks in a more limited way. For example, only attributes that appear in public datasets are used as quasi-identifiers. This method may cause a high risk of re-identifying the subject of personal information because of insufficient judgment on the additional background knowledge that the data recipient or attacker may have. Therefore, the process of identifying quasi-identifiers needs to consider both the characteristics of the data itself and the environment in which the data is used (application purpose, recipients, background knowledge, etc.).

C.2 Methods for identifying quasi-identifiers

The process of quasi-identifier identification begins after direct identifier identification. First, conduct preliminary identification on the characteristics of the data itself. Then analyze the environmental factors of data usage. Further screen the final quasi-identifiers.

a) Rapid identification of quasi-identifiers using prior knowledge: Candidate quasi-identifiers are quickly identified by comparing with recognized common quasi-identifiers. See Annex B for examples of common quasi-identifiers.

b) Further identification of quasi-identifiers via attribute correlation: Among the attributes of the target data set, the attributes with higher correlation are identified. For example, in the birth registration information database, the baby's date of birth and discharge date are highly correlated, and the date of birth is recognized as a common quasi-identifier, so the discharge date highly correlated with it is also usually identified as a quasi-identifier. Another example: There is also a high

correlation between medication and disease diagnosis, if any one of the attributes is identified as a quasi-identifier, the other is usually also identified as a quasi-identifier.

c) Screening of quasi-identifiers based on re-identification risk: The risk of re-identification of attribute values can be used to further screen quasi-identifiers. For each attribute, the uniqueness of its value can be calculated. The attribute with high uniqueness has a higher risk of re-identification. It is also possible to consider the impact of whether attributes are used as quasi-identifiers on the number of equivalence classes of the overall data set. For attributes that have a greater impact, for example, after being used as a quasi-identifier, the number of equivalence classes increases to a large extent relative to its inaction as a quasi-identifier, then the attribute needs to be considered as a quasi-identifier.

d) Screening of quasi-identifiers based on environmental risk: When determining the impact of environmental risks on the identification of quasi-identifiers, it is necessary to analyze from the perspective of the status and ability to obtain more background knowledge (background data) and the ability of data recipients to understand and analyze data.

1) Enterprises or institutions with more personal information, such as insurance companies (personal medical insurance), hospitals, e-commerce platforms, etc., usually have strong personal data acquisition capabilities. Therefore, the probability that such institutions use background knowledge for association re-identification is usually set as "high". For pharmaceutical or medical device companies, the background information of personal information they obtain may be very limited, so the possibility of re-identification can be set to "medium" or "low" (depending on the specific case requirements).

2) Data recipients with strong data understanding ability and analysis and processing ability have a higher risk of re-identification. On the contrary, if the knowledge and ability required for re-identification by using it exceed the knowledge and ability of the data receiver, the risk of re-identification is low.

3) Through the evaluation of environmental risks, attributes with low probability of re-identification using background information are usually not identified as quasi-identifiers. Those with high probability are usually identified as quasi-identifiers.

This is an excerpt of the PDF (Some pages are marked off intentionally)

------ The End ------