Personal data protection in information systems by de-identification and properties of de-identified data

K. F. Kerimov
Z. I. Azizova, email: z.i.azizova@mail.ru

Tashkent University of Information Technologies named after
Muhammad al-Khwarizmi

Аннотация. *This paper analyses and summarises the solutions used for de-identification. It also describes the properties of de-identified data and the requirements for such properties, characterises each property and illustrates the correspondence between de-identification methods and the properties of de-identified data.*

Ключевые слова: *personal data, de-identification, data anonymisation, quasi-identifiers, de-identified data, properties of de-identified data.*

Introduction

With the spread of information technology, organisations are becoming increasingly dependant on information systems and services and, as a result, more vulnerable to security threats. This has become apparent in the use of personal data processing information systems. The ever-accelerating informatisation of society and the rapid development of open information systems make data leakage and other forms of illegal access to the subjects' personal data much easier, which makes the task of ensuring its protection particularly significant and urgent.

1. Personal data protection methods and re-identification

De-identification is a process of detecting quasi-identifiers that directly or indirectly identify a subject (or object) and removing these identifiers from existing dataset. This process focuses on the inability to uniquely identify a subject based on the information or dataset. De-identified information is information that either originally contained no subject identifiers, or that has had such identifiers removed. There are 2 ways by which de-identification process can be done: safe harbour method and expert determination method [1]. The first one requires the elimination of all 18 personal identifiers. The last approach requires the preservation of certain personal identifiers (e.g. dates, demographic data, etc.) combined with an expert assurance that these identifiers cannot be used for re-identification. It should be noted that there is no strict definition of either the qualifications of the expert or the type of expertise, nor of the set of methods that produce the expert's opinion.

The importance of the de-identification process is due to the factors of availability of data sets that allow information to be used without compromising privacy. Protecting an individual or group from disclosure is another possibility for de-identification.

Re-identification of an de-identified dataset is made possible by the availability of high computing power of hardware and the widespread availability of publicly available information on the global network. This means that anonymised data can be linked to the specific entity to which it relates. Re-identification is implemented by combining two or more datasets to search the subject in both datasets. Such aggregation reveals information that directly identifies the subject. When an anonymised dataset is re-identified, both the direct and indirect identifiers are known and therefore the subject can be unambiguously identified.

The de-identification process minimises the risk of re-identification, but does not guarantee its impossibility. On the other hand, anonymisation techniques found in the literature, such as microdata aggregation, are almost always applicable only to tabular data and aim at guaranteeing a certain level of anonymity as a function of aggregation [2].

In the case of de-identification, the original information system of personal data is divided into subsystems: a subsystem with de-identification software (identification data), a de-identified data subsystem and a link element between the subsystems, one of them, except for the de-identified data, will have to be protected according to the requirements of the initial level or security class of the information system. Therefore, dividing the original system into subsystems in this way does not change the level of protection required for the system as a whole. The question arises as to the cost of organising the de-identification process and subsequent protection, which is a higher level than in the classical approach.

The method of de-identification of personal data must be appropriate to the purpose and objectives of data processing set by the data controller and must comply with the rules and requirements of the legal framework. The main purpose of any depersonalisation method is to ensure the anonymity of personal data. The main problem with most de-identification methods is the quasi-identifiers (indirect identifiers) that remain in the dataset of the de-identified data. They are identifiers that do not by themselves identify a person, but can be aggregated and linked to other information (publicly available data or other operators' database) to identify personal data subjects. The way in which personal data subjects are re-identified using this method is called a “link-based attack” [3].

2. Security management and basic personal data protection issues

The de-identification of personal data is performed in accordance with the methods and guidelines established by law. The result of the anonymisation process is that the personal data subject cannot be identified and that the de-identified data in question can be processed. This data, in its turn, must have a certain set of properties, as shown in Table 1.

Таблица 1

Description of the properties of the de-identified data

Name of property	Characteristic
Completeness	the need to retain the information about the personal data subject available prior to de identification making it impossible to uniquely identify personal data subjects after de identification, without the use of additional information.
Anonymity	the impossibility of unambiguously identifying the subject of personal data after de-identification without the use of additional information
Relevance	the possibility to process requests and to receive responses in the same semantic form when processing personal data
Structuring	the structural links between the de-identified data of personal data subjects are maintained and match the structural links before the de-identification
Semantic integrity	matching the semantics of the data when anonymised to the original dataset
Applicability	the possibility of processing personal data in an de-identified database of the personal data information system without prior de-identification of all records on the subjects

In addition to the properties of de-identified data and de-identification methods, there are mandatory requirements for the properties of the resulting de-identified data as shown in figure 1.

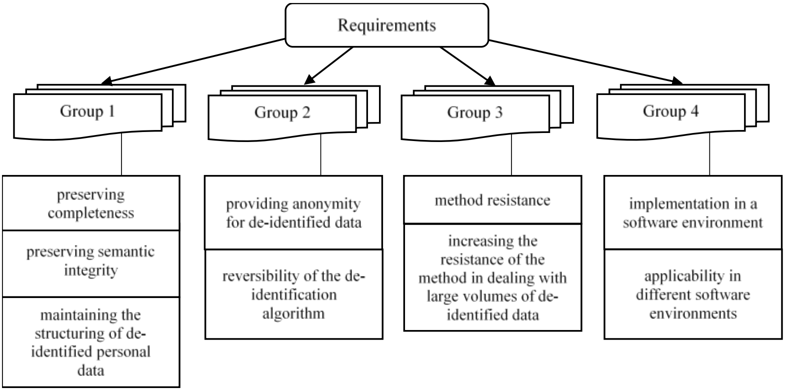
Requirements
Group 1
Group 2
Group 3
Group 4
preserving completeness
preserving semantic integrity
maintaining the structuring of de-identified personal data
providing anonymity for de-identified data
reversibility of the de-identification algorithm
method resistance
increasing the resistance of the method in dealing with large volumes of de-identified data
implementation in a software environment
applicability in different software environments

Рис. 1. Requirements for de-identified data

When processing de-identified data, one of the important tasks is to integrate the de-identification/re-identification procedure into the existing personal data processing systems, for example by providing an API application to the personal data controller with the possibility of forming anonymised database queries and implementing the de-identification procedure transparently for users. These requirements allow for the processing of anonymised data at minimum cost for upgrading existing data processing systems and maintaining user interfaces.

The scheme of correspondence between the properties of de-identified data and de-identification methods is shown in figure 2:

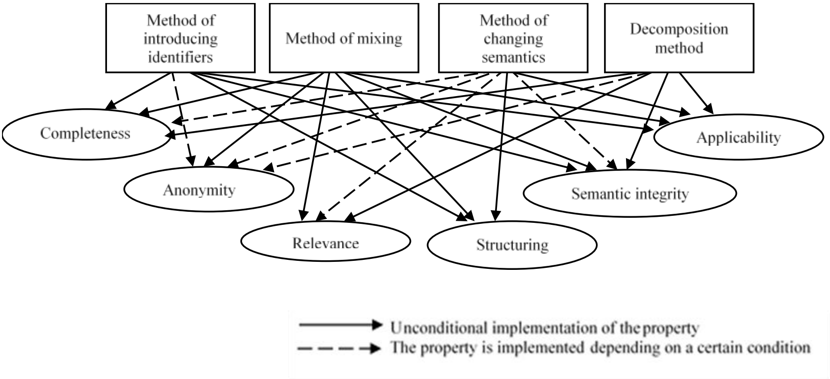
Method of introducing identifiers
Method of mixing
Method of changing semantics
Decomposition method
Completeness
Applicability
Anonymity
Semantic integrity
Relevance
Structuring
Unconditional implementation of the property
The property is implemented depending on a certain condition

Рис. 2. Correspondence between the properties of de-identified data and de-identification methods

Consequently, de-identified data loses the status of personal data. If in some personal data information system (or its segment) there is no access to re-identification software, de-identification algorithm, and it is supposed to work only with impersonal data, for example, for statistical conclusions or software development and debugging, such "external" database ceases to be an information system for personal data processing subject to the requirements of regulators.

Conclusion

As a trivial task, choosing the most appropriate depersonalisation method depends on the dataset, the extent to which the information is accessible to attackers and the type of information contained in the dataset in question. Because de-identified data loses its status as personal data, de-identified data can be transmitted through open communication channels without the need for additional security features. De-identification enables processing in distributed information systems that do not allow the personal data subject to be identified. In another case, if the database is stored in a de-identified form and re-identification occurs during processing, such personal data processing information system can be divided into segments, which in some cases may reduce the cost of protecting such a system as a whole by reducing the level of security of an individual segment.

References

1. Health Insurance Portability and Accountability Act (HIPAA). [Электронный ресурс] : база данных. – Режим доступа: https://www.cdc.gov/phlp/publications/topic/hipaa.html
2. An Investigation of the Consequences of Partial Aggregation of Micro-Economic Data /Feige, Edgar L & Watts, Harold W. // Econometrica, Econometric Society. – 2007. – vol. 40(2) – P. 343-360.
3. Anonymization of personal data. [Электронный ресурс] : база данных. – Режим доступа: https://lib.itsec.ru/articles2/focus/ob-obezlichivanii-personaljnyh-dannyh