Tashkent University of Information Technologies named after Muhammad al-Khwarizmi

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

K. F. Kerimov
Z. I. Azizova

Abstract:

In this article the problems of personal data protection by de-identification are discussed. The reasons for breach of confidentiality of personal data are discussed in detail, such as defining measures and means of personal data protection, minimising the costs of providing protection while complying with personal data protection requirements and other.

Keywords:

personal data, de-identification, data anonymisation, quasi-identifiers, de-identified data, properties of de-identified data.

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***.

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.

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 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 (Description of the properties of the de-identified data).

Table 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.

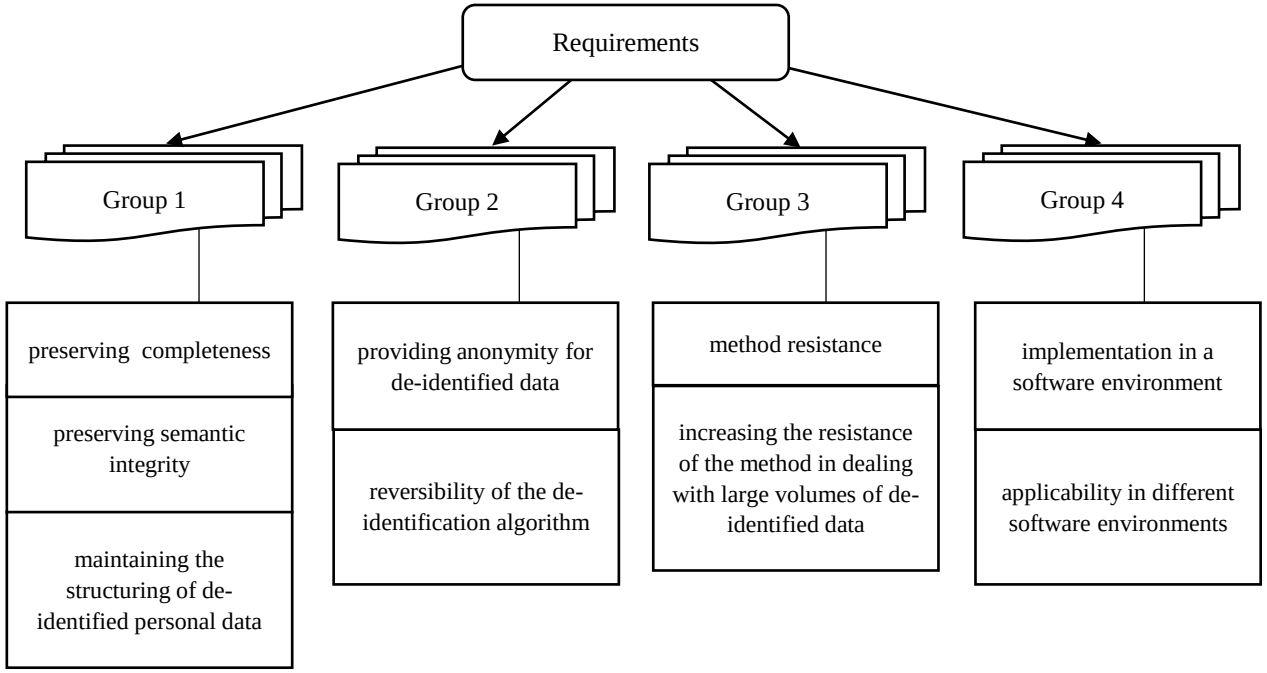

Figure 1. Requirements for de-identified data

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

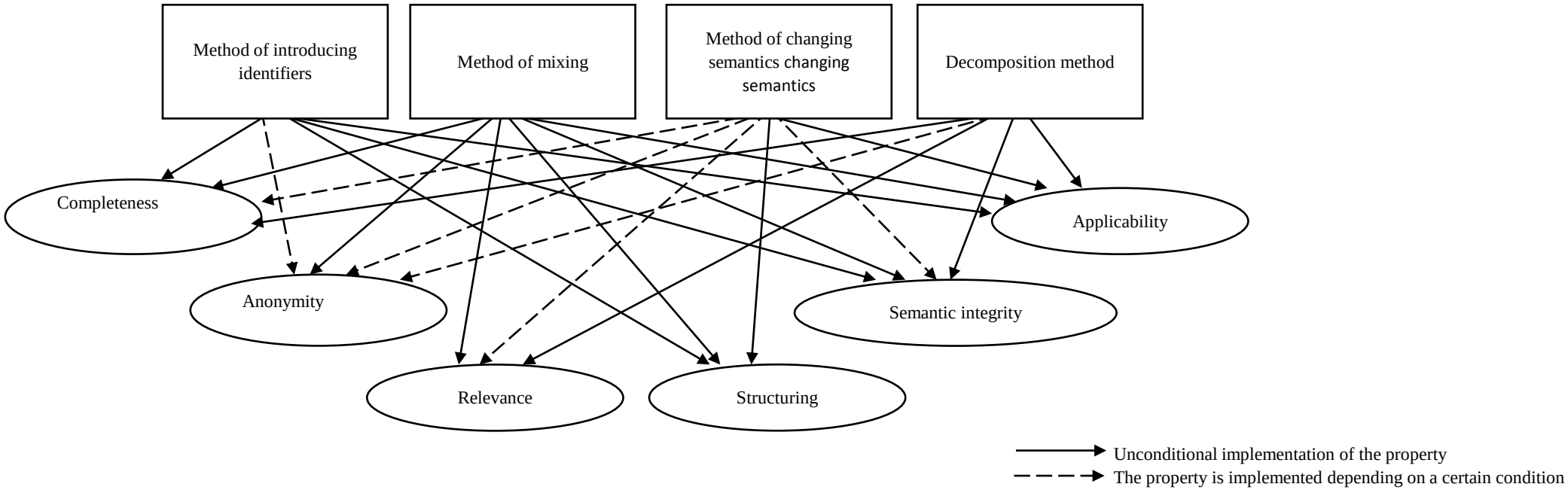

Figure 2. Correspondence between the properties of de-identified data and de-identification methods

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.

THANK YOU FOR ATTENTION!