

Protection of Personal Data Using Anonymization

Alexander V. Bogdanov[1], Alexander Degtyarev[1,4], Nadezhda Shchegoleva[1(✉)], Vladimir Korkhov[1,4], Valery Khvatov[2], Nodir Zaynalov[3], Jasur Kiyamov[1], Aleksandr Dik[1], and Anar Faradzhov[1]

[1] Saint Petersburg State University, St. Petersburg, Russia
{a.v.bogdanov,a.degtyarev,n.shchegoleva,v.korkhov}@spbu.ru,
{st080634,st087383,st069744}@student.spbu.ru
[2] DGT Technologies AG., Toronto, Canada
[3] Samarkand branch Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Samarkand, Uzbekistan
[4] Plekhanov Russian University of Economics, Moscow, Russia

Abstract. Data quality and security issues are very closely related. To ensure a high level of reliability in distributed systems and resilience from external attacks, the process of consolidating distributed data is critical. For consolidated systems, the access process relies heavily on data preprocessing, which, in turn, allows them to be anonymized. The analysis of closely related processes of consolidation and anonymization allows us to offer a secure access platform for distributed data, which makes it possible to implement secure access systems that depend only on the type and format of the data. It turns out that in the program stack for working with data, optimization can be done only with the entire framework, but not with its components. In this paper we perform analysis of data security as a complex problem related to both data quality and system architectures used to protect personal data.

Keywords: Big Data · Data access · Data anonymization · Data consolidation · Data platforms

1 Introduction

Recently, Big Data is increasingly turning from a fashionable area of research into a tool for practical use for many areas of activity. However, the increase in the volume and decrease in the connectivity of data leads to serious problems, both in terms of the appearance of security problems and possible data leaks. The latter is especially relevant when working with personal data and the emergence of a serious problem of data anonymization. The amount of information plays a key role in the complexity of the anonymization process. The emergence of new sources of information makes it possible to compare them with those published (available) earlier and inevitably leads to the emergence of risks of re-identification of previously depersonalized data. This, in turn, forces us to abandon the concept of guaranteed anonymized data, introducing certain

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12956, pp. 447–459, 2021.
https://doi.org/10.1007/978-3-030-87010-2_33

boundaries for dealing with risks and building a continuous threat assessment process. In doing so, it is important to consider two critical parameters that determine data security: data type and system connectivity [1].

To clarify this circumstance, we will consider three types of systems - a closed system, a private cloud, and a public cloud. For a closed system, these problems do not arise and there are a large number of products that solve this problem for both scientific and commercial structures. For private clouds, the situation becomes more complicated and a number of dangers appear. To ensure security, it is necessary to introduce a two-level access system, similar to the GRID system, and, in addition, it is necessary to consolidate data, and the consolidation is multi-level, not all stages of which are currently optimized [2].

For public clouds, these issues are on the rise, with data quality and criteria for selecting anonymization models being critical issues. The issue of data quality has been investigated in sufficient detail, and the conclusion from this study is very tough - before any work with data, the data quality must be checked and all possible adjustments must be implemented. If the data is well prepared, you can start data pre-processing.

The main stages of preparation are the following:

1. Anonymization:
   - Anonymization of personal data must be built into the overall data management system (Privacy By Design)
   - The results of anonymization should be measurable (benefits and risks)
   - Anonymization as a measure of information protection should include not only technical, but also organizational measures
   - Data should be categorized according to the degree of sensitivity, for different categories different methods should be used
2. Data quality check:
   - completeness (preservation of all information about specific subjects or groups of subjects that was available before depersonalization);
   - structuredness (preservation of structural links between the impersonal data of a specific subject or a group of subjects corresponding to the connections that existed before anonymization);
   - semantic integrity (preservation of the semantics of personal data during their depersonalization);
   - anonymity (impossibility of unambiguous identification of data subjects obtained as a result of anonymization);
3. Data Consolidation:
   - coordination of metamodels;
   - interpretation of data models;
   - comparison of classifiers, directories and domains;
   - combining information.

When implementing this process, a number of problems arise:

- conceptual differences between data models used in different sources within the same operating environment;

– interpretation of data models;
– comparison of classifiers, directories and domains;
– combining information.

The procedural approach implies hard coding of integration procedures based on specific data models. The advantages of this approach include the fact that there are practically no restrictions on the use of different structures and models in the database. The disadvantages are the following:

– for each interaction of data sources, you need to implement your own software integration module that uses the full cycle of operations;
– complexity of maintenance;
– hindered efficiency;
– instability of the software to change the data model.

There is no standard set of tools for solving these problems yet, so at this stage it is important to build a set of platform requirements and see on what basis such a solution can be built [3].

2 Ecosystems to Process and Store Super Large Amounts of Data

The architecture of the Big Data processing system uses complex systems in which several components or layers can be distinguished. Usually, there are four levels of components of such systems: reception, collection, analysis of data and presentation of results [2]. This division is largely arbitrary, since, on the one hand, each component can be divided into subcomponents, and on the other hand, some functions of the components can be redistributed depending on the problem being solved and the software used, for example, they allocate data storage in a separate layer.

To work with Big Data, system developers create data models that are related to the real world. Developing adequate data models is a complex analytical task performed by system architects and analysts. The data model allows you to create a mathematical model of interactions between objects in the real world and includes a description of the data structure, data manipulation methods and aspects of maintaining data integrity. Describing the development of data models is beyond the scope of this work. Distributed systems of various types are used to store data. These can be file systems, databases, journals, mechanisms for accessing shared virtual memory. Most storage systems focused exclusively on working with Big Data have an extremely limited number of functions due to the inherent complexity of creating highly efficient distributed systems.

In order to work with data faster, data storage and processing systems are parallelized in a computing cluster. However, according to Brewer's hypothesis, it is impossible to ensure the simultaneous data consistency, data availability and system resistance to the separation of individual nodes [7]. The hypothesis is proven for transactions of the ACID type (Atomic, Consistent. Isolated, Durable) and is known as the CAP theorem (Consistency, Availability, Partition tolerance).

2.1 Data Ingestion

Data sources have various parameters, such as the frequency of data receipt from the source, the size of the data portion, the data transfer rate, the type of data received and their reliability. To efficiently collect data, you need to establish data sources. These can be data warehouses, aggregated data providers, API of any sensors, system logs, human-generated content in social networks, in corporate information systems, geophysical information, scientific information, inherited data from other systems [4]. Data sources define the original data format. For example, we can independently conduct weather research on the territory of the airport, use data from aircraft taking off and landing, purchase data from satellites flying over airports from the local weather service, and also find it somewhere else on the network. In general, for each source, it is necessary to create its own collector (Data Crawler for collecting information on the network and Data Acquisition for making measurements). Receiving data consists in the initial preparation of data from sources in order to bring the data to a common data presentation format. This uniform format is chosen according to the accepted data model. Conversions of measurement systems, types (typification), verification are performed. Data processing does not meaningfully affect the information available in the data, but it can change its representation (for example, bring coordinates to a single coordinate system, and values to a single dimension).

2.2 Data Staging

The data collection phase is characterized by direct interaction with data storage systems. A collection point is established at which the collected data is supplied with local metadata and placed in a repository or transmitted for further processing. Data that for some reason did not pass the collection point is ignored.

For structured data, conversion from the original format is carried out according to predefined algorithms. This is the most efficient procedure when the data structure is known. However, if the data is presented in binary form, the structure and connections between the data are lost, then the development of algorithms and software based on them for data processing can be extremely difficult [2].

Semi-structured data requires the interpretation of the incoming data and the use of software that can work with the used data description language. A significant advantage of semi-structured data is that it often contains not only the data itself, but metadata in the form of information about the relationships between data and how to obtain it. The development of software for processing semi-structured data is a rather complex task. However, there are a significant number of ready-made converters that can, for example, extract data from XML format into a generated table view.

The largest amount of effort is required for processing unstructured data. To translate them into a given format may require the creation of special software, complex manual processing, recognition and selective manual control.

During the collection phase, data types are checked and basic data validation can be performed. For example, the coordinates of gas molecules contained in any region cannot lie outside this region, and the velocities cannot significantly exceed the speed of sound. In order to avoid typing errors, it is necessary to check if the units of measurement are specified correctly. For example, one dataset might measure elevation in kilometers and another in feet. In this case, it is necessary to convert the height to those units of measurement that are accepted in the used model.

When collected, data is systematized and provided with metadata stored in associated metadata. If you have a large number of data sources, you may need to manage the collection of data in order to balance the amount of information coming from different sources.

2.3 Analysis Layer

Data analysis, unlike data collection, uses the information contained in the data itself. The analysis can be carried out both in real time and in batch mode. Data analysis is the main task in terms of labor intensity when working with Big Data. There are many data processing techniques: predictive analysis, queries and reporting, mathematical model reconstruction, translation, analytical processing, and others. The techniques use specific algorithms depending on the goals set. For example, analytical processing can be image analysis, social media analysis, geographic location, feature recognition, text analysis, statistical processing, voice analysis, transcription. Data analysis algorithms, like data processing algorithms, rely on a data model. In this case, in the analysis, several models can be used that set the general data format, but simulate the meaningful processes in different ways, the data about which we process. When using artificial intelligence methods, in particular neural networks, in the analysis, the models are dynamically trained on various data sets.

When analyzing data, the entities described by the data are identified based on the information available in the data and the models used. The essence of analysis is an analytical mechanism that uses analytical algorithms, model management and entity identification to obtain new meaningful information that is the result of the analysis.

2.4 Consumption Layer

Data analysis results are provided at the consumption level. There are several mechanisms to use the results of Big Data analysis.

- Monitoring metadata.
  Subsystem for real-time display of essential parameters of the system, workload of computers, distribution of tasks in a cluster, distribution of information in storages, availability of free space in storages, data flow from sources, user activity, equipment failures, etc.

- Data monitoring.
  Subsystem for displaying in real time the processes of receiving, collecting and analyzing data, data navigation.
- Generation of reports, queries to data, presentation of data in the form of visualization on dashboards (Dashboard), in PDF format, infographics, pivot tables and quick references.
- Data transformation and export to other systems, interface with BI systems

3 Anonymizing the Data

In this paper, we consider the classical methods of anonymization, which are recommended in the standards, such as ISO 20889:2018 and NIST 8062. New modern methods are emerging, this is separate topic with some important methodological achievements available [7].

3.1 Approaches to Anonymization

Basic Methods of Anonymization

- Introduction of quasi-identifiers [8]. Methods of depersonalization based on the introduction of replacement identifiers - replacing part of the information attributes of personal data. The need to create conversion tables.
- Aggregation of data [8]. Aggregation - as part of the aggregation of data and thereby deletion of personal data.
- Data decomposition [8]. Separation of sets of personal data in order to protect personal identifiers. Organization of separate storage and processing of data. Boundary calculations.

Statistical Methods

- Sampling. Construction of approximate data, based on statistical data processing.
- Aggregation [11]. Formation of generalized data marts.

Cryptographic Methods

- Deterministic Encryption. Create unidirectional text data transformation using public keys - RSA Based.
- Encryption preserving the order of the attribute or format. Order-preserving encrypted, Format–preserving encryption - The order-preserving encryption scheme (OPES) allows you to directly apply comparison operations to encrypted data without decrypting the operands, so aggregated data can be applied directly to encrypted data.
- Homomorphic encryption [12]. Encryption of personal data while maintaining the ability to analyze and search for hidden data - RSA, El Gamal cryptosystem.
- Quantum resistance. The resistance of encryption to quantum computing.

Attribute Suppression Techniques

- Local suppression. Method of replacing several values of sensitive attributes with a missing value, can lead to data loss.
- Suppression of records. Record Suppression. Suppression of a number of records relating to one subject of personal data.

Pseudo-anonymization Methods

- Replacement of attributes. Selection Of Attributes - replacement of attributes from the number of substitution dictionaries.
- Use of aliases. Anonymization methods based on the introduction of the replacement of identifiers - replacement of part of the information attributes of personal data.

Summarizing Data

- Output of approximate values. Rounding – output of approximate values of attributes of personal data subjects.
- Method of limit values. Top and Bottom Coding - setting the upper and lower limits, exceeding or underestimation of which leads to the output of the limits themselves.
- Complex Attributes. Set Of Attributes-setting a combination of attributes instead of individual attributes.

Randomization

- Adding noise. Noise Addition, Blurring-Embedding of random data in a given set, distortion of sensitive attributes.
- Microaggregation [10]. Micro Aggregation is a special clustering task whose purpose is to group a set of points into groups of at least k points so that the groups are as homogeneous as possible.
- Shuffling of attributes [9]. Permutation, Swapping - entanglement of attributes between several subjects.

Privacy Models (Anonymization Metrics) – Protecting a dataset from re-identification by setting thresholds for the proportion of records. To do this, you must specify the basic information, based on this data, statistical models are used to evaluate the characteristics.

k-anonymity. The database is a table with n rows and m columns. Each row in a table is a record for a specific member of the population, and records in different rows do not have to be unique. The values in the various columns are the values of the attributes associated with the members of the population.

l-diversity. Given the existence of such attacks, in which sensitive attributes can be inferred for k-anonymity data, the l-diversity method was created to increase k-anonymity by further preserving the diversity of sensitive fields.

t-closeness. It is a further refinement of l-diversity group anonymization, which is used to preserve confidentiality in datasets by reducing the granularity of data

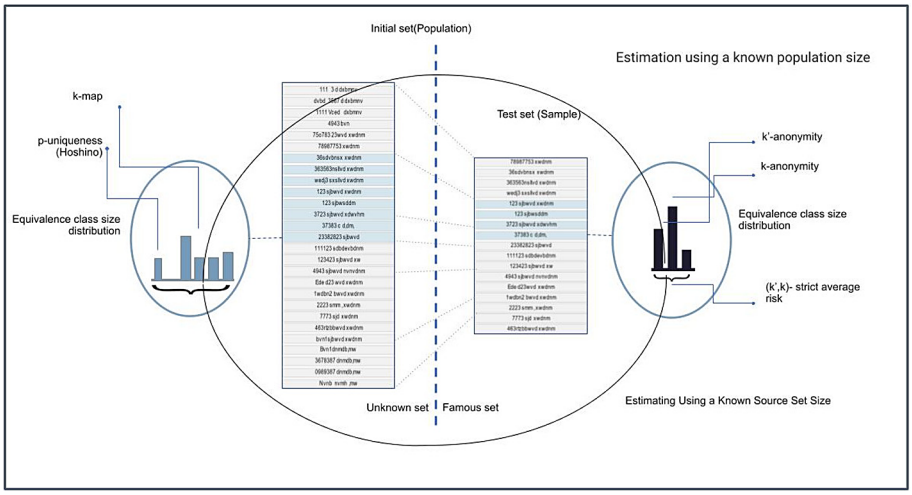

Fig. 1. k-anonymity illustration

presentation. This reduction is a tradeoff that results in some loss of efficiency in data management or data mining algorithms to achieve some privacy. The t-closeness model extends the l-diversity model by treating attribute values separately, taking into account the distribution of data values for that attribute (Fig. 1).

The Differential Privacy Model. [13] is a system for publicly exchanging information about a dataset by describing patterns of groups in the dataset without disclosing information about the individuals in the dataset. The idea behind differential privacy is that if the effect of doing an arbitrary single replacement on a database is small enough, the result of the query cannot be used to infer much about any particular person and therefore provides privacy. Another way of describing differential confidentiality is by limiting the algorithms used to publish aggregate information about the statistical database, which limits the disclosure of private information about the records that are in the database. For example, differentially private algorithms are used by some government agencies to publish demographic information or other statistical aggregates while maintaining the confidentiality of survey responses, and by companies to collect information about user behavior while monitoring what is visible even to internal analysts. Roughly speaking, an algorithm is differentially private if the observer, seeing its output, cannot say whether the information of a particular person was used in the calculations. Differential confidentiality is often discussed in the context of identifying individuals who may be in the database. While this does not directly apply to identification and re-identification attacks, differentially private algorithms are likely to resist such attacks (Table 1).

Table 1. Example of implementation and features of calculating risk

Anonymization method	Application	Risk calculation features
Permutation, Swapping	For the selected general group, the change of quasi-identifiers	Requires extensive attribute processing
Local Generalization	example: address →street → city →region	Based on anonymization metrics
Aggregation	Representation of generalized records	For big data and rare groups, the risk of pooled sets
Reduction	Used as a last resort as it destroys data	Significant degradation of data quality
Local Suppression	For example, to hide rare diseases, or the family status of young participants	Deterioration in utility. Risk - according to anonymization metrics
Masking	For example, credit cards are represented as XXXX XXXX XXXX6789	Loss of data quality, only suitable for certain data types
Pseudo-anonymization	For example, by splitting data or direct hashing	Vulnerable by preserving data uniqueness
Deterministic encryption	Create searchable ciphertexts	Limited to use on relatively small datasets
Homomorphic encryption	Creation of encrypted and searchable texts	Limited to use over relatively small datasets
Sampling	Small set statistical construction	Loss of quality due to safety

3.2 Privacy Models

Risk is a possible damage, that is, a combination (as a rule, a product), the likelihood of a threat being realized and damage from it.

Note that the threat and risk are determined not in general, but in relation to a specific protected resource. In the terminology of the theory of business process management, instead of a resource, a synonymous concept is used - an asset, which defines everything that has value for an organization, for example: databases, hardware and software, network infrastructure, specialists.

In the process of depersonalization, not only information about identifiers is lost, but also links between individual records, dependencies, information about average values, etc. It is difficult to estimate such losses on a sufficiently large data set, therefore, utility measures are applied:

1. Classification metric - loss of data value when performing a classification task;
2. Reuse metric - loss of granularity (on the number of unique attribute values within an equivalence class);

3. Shannon's measure is a measure based on the entropy of information, based on the assumption that the increase in information is equal to the lost uncertainty;
4. The measure of mutual utility is a measure based on the assessment of the granulation of information; it decreases with an increase in the size of the equivalence class (Fig. 2).

Graph Metrics	Pre-Analysis	Iteration-1	Iteration-2	Iteration-3
Nodes	36692	39521	40496	40983
Edges	183831	193726	205819	207883
Degree Distribution	5.01	4.9	5.08	5.07
Diameter	11	12	13	14
Clustering Coefficient	0.49	0.51	0.53	0.56
Data Utility		96.34	90.75	85.75

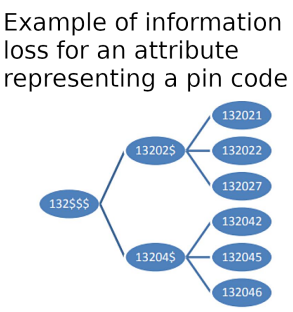

Fig. 2. Metrics of usefulness

Information security threats are classified according to a number of criteria:

– due to the occurrence (natural or man-made, including intentional or accidental);
– by source location (external or internal);
– by compromising subsystem or segment (network, cryptographic, etc.);
– by the stage of formation in the life cycle of the system (implementation and operational); by the resulting action (violate the level of confidentiality, integrity or availability).

3.3 Privacy Testing

Scripting Method. This method is the most natural in risk analysis, since it involves the formulation of those scenarios that can be realized in practice, and the assessment of financial results in each of these scenarios. In a particular case, when at each moment of time each of the recognized significant factors takes on a finite number of values (or the values are combined into a finite number of distinguishable groups), a scenario tree is drawn up. However, a direct analysis of all the possibilities is often impossible. Therefore, in most cases, some criteria are developed by which the scenarios are compared, and only a part of the scenarios called non-dominated is analyzed. In this formulation, the scenario method is somewhat modified - not those scenarios that are realized with different values of the uncertainty factors are considered, instead of them, the sets of our criteria

are considered, depending on the decision we have already made, and the risk here can be taken into account as one of the criteria.

An obvious example of such criteria can be profitability (which must be maximized) and risk (it must be minimized). It is hardly possible to obtain a solution that is optimal for both criteria. It is clear, however, that a scenario in which both the risk is greater and the return is less always yields to the scenario with a higher return and less risk. Thus, among the scenarios, one can single out those for which there are, of course, the best (dominant) scenarios, and the remaining - non-dominant, which should be used as candidates for choosing a rational solution.

Decision-making on a large number of sometimes conflicting criteria from the point of view of a mathematician is one of the problems of multicriteria optimization. Another typical object arising in the formalization of the problem of financial risk management is the problem of stochastic optimization, which takes into account the presence of a random component. In this case, in the presence of randomness, the average value of some indicator is optimized. One of the numerical methods for solving the stochastic optimization problem is the Monte Carlo method (Fig. 3).

Monte Carlo Method. (one of the simulation methods) is applied when it is difficult to analytically calculate the mathematical expectation of a particular indicator. One way to calculate averages (as well as probabilities) is to run a large number of experiments. A series of random numbers generated by the corresponding sensor (more precisely, they should be called pseudo-random) are used as data for the assessment. Based on these data, a large number of scenarios are simulated, and to estimate the unknown mathematical expectation in accordance with the law of large numbers, the sample mean is used, which is the desired numerical estimate.

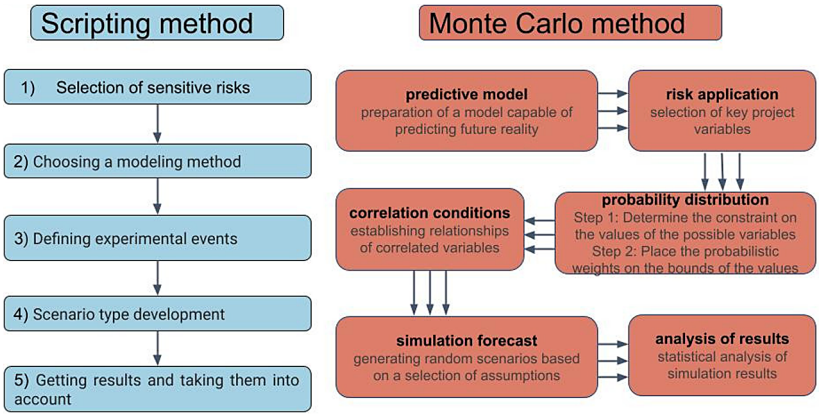

Fig. 3. Stress testing algorithm

In financial risk management, problems requiring the use of the Monte Carlo method arise quite often. It is important to understand that using this method involves a lot of computation. Therefore, if the system uses this method to help traders make decisions in real time, it requires serious computing power. Especially often the Monte Carlo method is used to solve the problem of pricing the so-called exotic options.

4 Conclusion

The issue of data security is a complex problem closely related to both the quality of data and the architecture of the systems used to process and store them. It becomes especially acute with the transition to Big Data. To build such systems, you have to apply completely new approaches to security, essentially determined by the type of data. As a result of the application of anonymization methods, data are obtained that have varying degrees of anonymization. The security of the dataset obtained cannot be judged solely on the basis of the data itself. External and internal connections have a significant impact: the availability of additional information, the attacker's motivation, the legal framework, the IT systems used, management practices, etc. This leads to the division of risk between the risks of the data itself (taking into account the applied methods of anonymization) and the risks of the environment (contextual risks). Most are based on the concept of an equivalence class - the ability to select identical records within a dataset in terms of quasi-identifiers. Therefore, the main parameters for evaluating the model depend on the type and quality of data, and not on a specific case, since for different types of attacks, attackers (such as a "marketer", "prosecutor", "journalist", which can be used in the scenario method) are based on the size of the equivalence class of the data itself.

References

1. Bogdanov, A., Uteshev, A., Khvatov, V.: Error detection in the decentralized voting protocol. In: Misra, S., et al. (eds.) ICCSA 2019. LNCS, vol. 11620, pp. 485–494. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-24296-1_38
2. Bogdanov, A., Degtyarev, A., Shchegoleva, N., Khvatov, V.: Data quality in a decentralized environment. In: Gervasi, O., et al. (eds.) ICCSA 2020. LNCS, vol. 12251, pp. 58–71. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58808-3_6
3. Bogdanov, A.: Private cloud vs personal supercomputer. Distributed Computing and GRID Technologies in Science and Education, pp. 57–59. JINR, Dubna (2012)
4. Bogdanov, A, Shchegoleva, N., Ulitina, I.: Database ecosystem is the way to data lakes. In: Proceedings of the 27th Symposium on Nuclear Electronics and Computing (NEC 2019)
5. Rinku, N., et al.: Detection of forgery in art paintings using machine learning. Int, J. Innov. Res. Sci. Eng. Technol. **6**(5), 8681–8692 (2017)
6. Di Pietro, R., Salleras, X., Signorini, M., Waisbard, E.: A blockchain-based distributed Trust System for the Internet of Things (2018)

7. The Trust Chain Consensus, COTI Whitepaper. https://coti.io/files/COTI-technical-whitepaper.pdf
8. Guidelines for the application of Roskomnadzor order of September 5, 2013 N 996 "On approval of requirements and methods for anonymization of personal data"
9. The Anonymization Decision-Making Framework, UKAN (2016)
10. El Emam, K., Arbuckle, L.: Anonymizing Health Data. O'Reilly, Sebastopol (2014)
11. Tomashchuk, O., Van Landuyt, D., Pletea, D., Wuyts, K., Joosen, W.: A data utility-driven benchmark for de-identification methods. In: Gritzalis, S., Weippl, E.R., Katsikas, S.K., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) TrustBus 2019. LNCS, vol. 11711, pp. 63–77. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27813-7_5
12. Iyengar, V.S.: Transforming data to satisfy privacy constraints. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2002, p. 279 (2002)
13. NISTIR 8062. An Introduction to Privacy Engineering and Risk Management in Federal Systems
14. Ito, K., Kogure, J., Shimoyama, T., Tsuda, H.: De-identification and Encryption Technologies to Protect Personal Information (2016)
15. Gómez-Silva, M.J., Armingol, J.M., de la Escalera, A.: Triplet permutation method for deep learning of single-shot person re-identification. Fujitsu Sci. Tech. J. **52**(3), 28–36 (2020)
16. Lee, Y.J., Lee, K.H.: What are the optimum quasi-identifiers to reidentify medical records? In: ICACT Transactions on Advanced Communications Technology (TACT), vol. 6, no. 4, July 2017