Theoretical aspects of the protection of personal data of employees of the enterprise by the method of pseudonymization

Andrey Gazizov[1,*], *Evgeny* Gazizov[2], and *Svetlana* Gazizova[3]

[1]Don State Technical University, 344003, Gagarin sq., 1, Rostov on Don, Russia
[2]Kazan state agrarian University, 420015, 65 K. Marx str., Kazan, Russia
[3]Kazan (Volga region) Federal University, 420008, 18 Kremlevskaya str., Kazan, Russia

Abstract. The topic of pseudonymization of personal data has shown, that theoretical and methodological basics in sphere of automatized systems have just started to gain general trend. The majority of studies in this sphere are, commonly, about personal data in general, rarely touching the topic of pseudonymization and depersonalization. Therefore, the topic of pseudonymization has not fully assimilated in enterprise systems and has not grown any popularity, because enterprises tend to choose reliable tools and methods of information security while depersonalization is only beginning its way and is not common for big corporations. This leads to disinterestedness in solving known issues and goals of pseudonymization, universal methods have not been researched. However, low cost and simplicity of this method of personal data protection is turning our attention on it and ask ourselves a question: “Should we have a deep dive in it?”. Answer is obvious – yes. Certainly, this method has its disadvantages and it is not an ideal solution. But it certainly should be distributed worldwide.

1 Content of the concept of pseudonymization

When the Federal law "on personal data" was approved on July 27, 2006, it became clear that the protection of personal data in information systems has reached a new level of efficiency, given the increased amount of storage and processing of necessary data for regulating socially important mechanisms, both at the governmental and public level. The growth rate of information technologies currently allows users of various categories to access multifunctional data banks.

At the same time," on the other side " of Informatization, hackers who want to get access to classified data are finding more and more new and sophisticated ways to achieve their own goals, so the tasks of ensuring the protection of such data play a crucial role, both at the local Russian and global level.

Personal data information system is a set of personal data stored in databases that allow their processing using information and communication technologies.

[*] Corresponding author: gazandre@yandex.ru

Thus, the protection of personal data is carried out by eliminating illegal and unauthorized access to them, in order to avoid their declassification, modification, elimination and other illegal actions. For this purpose, organizational and technical measures are implemented in personal data information systems to protect information from leakage using hardware and software, including those that depend on the human factor. Such personal data protection measures should provide protection:

1) Data representing electrical signals.
2) audio recordings of negotiations.
3) Data representing physical fields.
4) Data stored on data carriers.

One of the most reliable ways to protect personal data is pseudonymization, i.e. a method of depersonalization of data that hides the connection with the data subject and creates a link between a set of characteristics of the subject and its aliases (fictitious names), if there are several of them[1].

Pseudonymization is essentially a special case of depersonalization and de-identification (the process of removing the connection between identifying data and their subject); where an alias is just an identifier that serves as a reference to the original data of the data subject, which is usually stored in different segments of the information system.

There are two types of pseudonymization:

- reversible;
- irreversible.

The difference between them is that the irreversible pseudonymization method completely removes any connection between false and true data; therefore, their recovery is no longer possible.

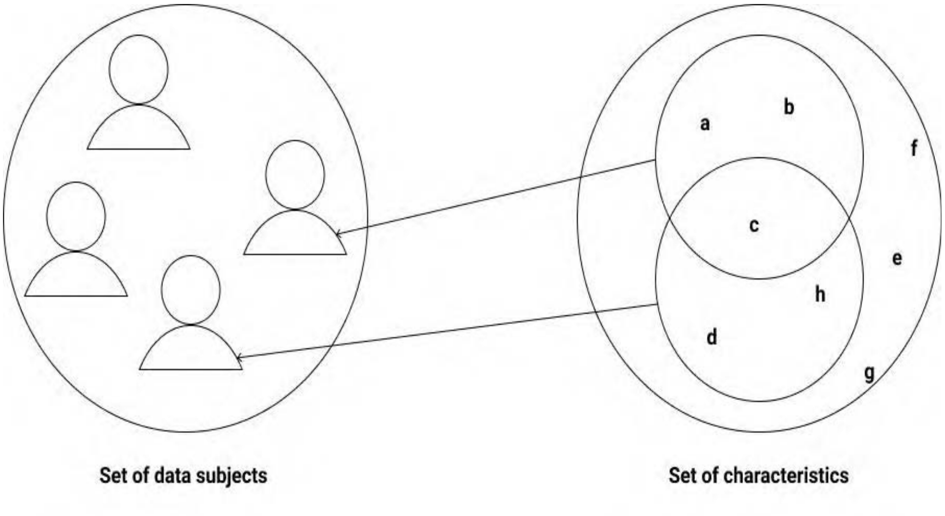

Fig 1. Method for identifying data subjects

In 2016, the European data protection standard GDPR (General data protection regulation) was released, which describes personal data as "any information that somehow relates to an identified or identifiable individual", so the most important advantage of pseudonymization is the ability to aggregate (Association) data for persons who have access to this information, but it does not violate the rules for access to confidential information of data subjects [2].

According figure 1 several data subjects that have a number of characteristics (address, first name, last name, etc.) stored in the company's database and related to the personal data of these subjects.

The subject of personal data is determined on the basis of a set of subjects among which it can be identified; thus, it will be quite easy to identify the identified data subject by a number of characteristics. Often a single unique characteristic is sufficient for this purpose; in other cases, a number of descriptions have to be used to reduce the search for the data subject. However, some characteristics may have properties of impermanence, that is, they may change over time (for example, the address of residence or, more often, the email address)[3].

2 Types of personal data

Personal data is divided into two parts according to the criteria for their identification:
- processed, i.e. data that describes characteristics that make it impossible to identify the data subject;
- identifying data, i.e. data that describes characteristics by which it is possible to identify the data subject.

However, there are conflicting cases when identifying data is also processed, so in an effort to develop a method for pseudonymizing personal data, it is necessary to take into account the "aggregation" of identifying data in order to reduce their level. In some cases, this is not possible, so such identification risks must be specified in the company's security policy [4].

3 The concept of pseudonymization

The availability of personal data for quality control and subsequent improvement by eliminating deficiencies in the enterprise information system plays an important role in the development of the enterprise. However, according to the confidentiality requirements, this data must be changed in such a way as to hide the identity of the subject.

No security method can guarantee the absolute result of data concealment, while complying with all the requirements and norms established by standards in the Russian Federation. Therefore, for such cases, a model of personal data leakage threats should be developed that would take into account the following factors:
- determine the purpose of personal data processing;
- minimize the information that is provided to achieve the goal of personal data processing;
- take into account the threat of personal data leakage;
- comply with the rules for accessing personal data.

Ways to classify identities, access to data, and threats are determined based on the above factors. This iteration should be performed with each new access to personal data, despite the fact that different processes may require a common strategy for the actions taken. In most cases, the protection of personal data is based on the General goals and standards to the minimum necessary protection of information. In other words, the personal data protection strategy will be common, but the methods of its implementation will differ[3].

Data depersonalization is widespread (figure 3), i.e. the process of destroying the connection between a set of characteristics and a personal data subject without being able to restore it. It is achieved in two ways:

– by completely deleting or changing the characteristics, in consequence of which the link is lost or ceases to be unique, pointing to several subjects at once;

– by increasing the number of data subjects, so that the link is no longer unique and points to multiple data.

At the same time, pseudonymization (figure 4) also removes the relationship between the characteristic set and the data subject; however, it adds a relationship between characteristics and aliases. As a result, it is quite possible to establish a link between some pseudo-named data that relates to a particular subject, without declassifying the identity.

Pseudonymization can be reversible or irreversible. Irreversible pseudonymization excludes the possibility of identifying a data subject by its set of characteristics and pseudonyms; reversible pseudonymization, on the contrary, includes this possibility. You can do this in the following ways:
1) Using the processed characteristics (by decrypting the identifying information that is stored in the processed characteristics).
2) Using an alias (matching table).

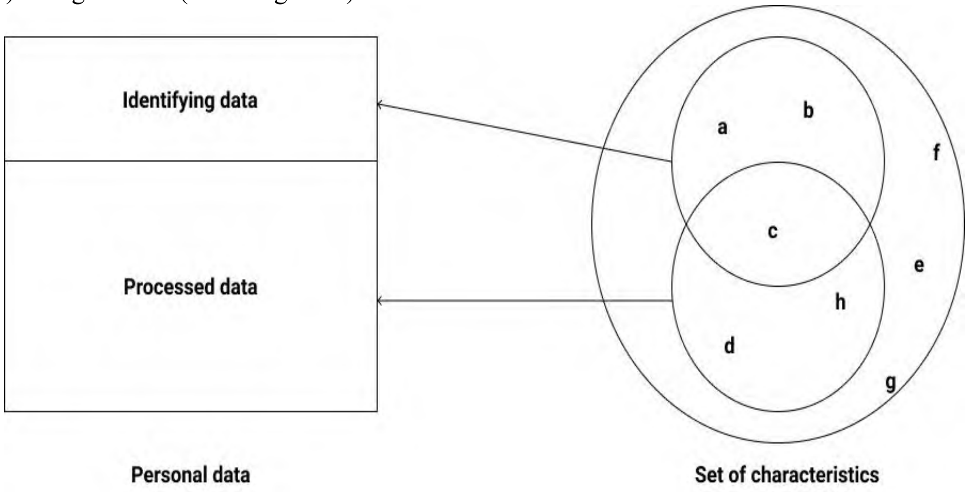

Fig 2. Types of personal data

There are several ways to implement reversible pseudonymization; however, it is worth noting that the procedure for converting pseudonyms to identifying data must be performed by individuals or businesses that have received permission to do so; because it requires a much higher level of protection than irreversible pseudonymization.

Pseudonymized data is grouped according to certain criteria, which are defined in the descriptions of personal data characteristics, from which the pseudonymized data is directly derived [4,5].

4 Classification of aliases depending on threats and channels of information leakage

Just as when building any information system, you must take into account the degree of threats and ways of information leakage, so for pseudonymization, you must take into account the probability of data identification and understand the interaction of entities and relationships in a particular database. In some cases, you need to take into account possible ways to accidentally declassify information; and in others-to anticipate and prevent unauthorized actions of hackers, which is defined and established in the company's security policy[6].

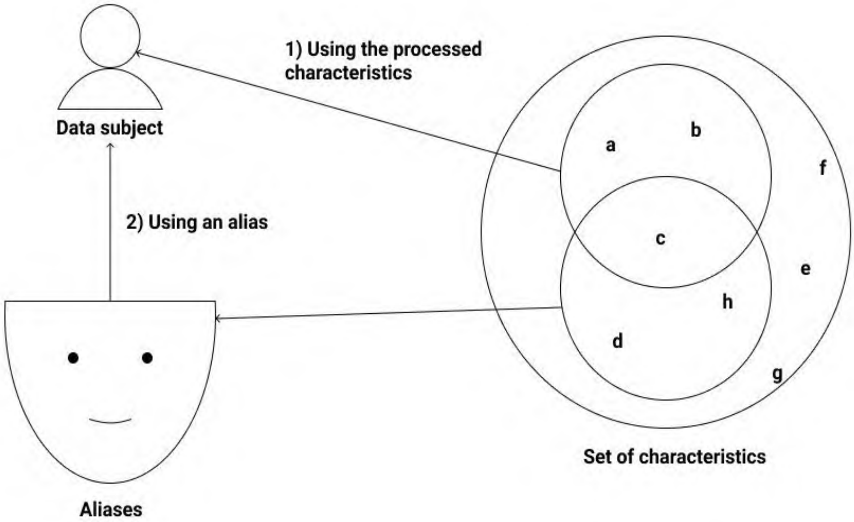

Fig 3. The method of pseudonymization

The probability of information leakage exists regardless of how any particular personal data protection algorithm handles its functions. Therefore, it is necessary to predict all possible channels of leakage of this data and analyze the results of a de-identified set of characteristics of any groups that are necessary to restore a classified identity. This approach is critical if certain identifying characteristics are required for processing purposes.

When creating threat models, it is important to consider not only direct identifying data, but also indirect data that affects several data subjects at once. It is necessary to predict every scenario of requirements of the accepted identifiers. Based on this, determine which personal data to fill in with null values, which to convert, which to leave intact, and instead of which to enter a pseudonym. At the same time, it is possible to use three levels of the pseudonymization procedure, which provide the necessary level of personal data protection (table 1). They are based on threats to restore the original data of the subject, taking into account direct and indirect identifying data [7].

At the same time, the assessment of the degree of threats to the restoration of the subject's identity must be made repeatedly at all levels, as indicated in the company's security policy, and reviewed several times a year.

Table 1. Levels of the pseudonymization procedure

№ level's	Threats related to it
1	Applying data elements that identify the subject
2	Search among shared data

3	Abnormal information in the database

Table 2. Security Measures depending on the level of confidentiality

№ level's	Possible data leakage channels	Ways to protect confidential data
1	Accidental declassification, the human factor	Removing obvious or easily accessible indirect data that instantly reveals the identity of the subject
2	Accidental declassification, human factor, technical means of unauthorized access by an attacker	Removing obvious or easily accessible indirect data that instantly reveals the identity of the subject, analyzing threats to reveal the identity of the subject, removing absolute timestamps in the database structure
3	Accidental declassification, human factor, technical means of unauthorized access by an attacker, access to abnormal data	Removing obvious or easily accessible indirect data that instantly reveals the identity of the subject, analyzing threats to reveal the identity of the subject, removing absolute timestamps in the database structure

In addition to regularly established threat assessment procedures, it is possible to review them in connection with a certain event (for example, making large and large-scale changes to the database).

Depending on the level, the number of checks may also change; however, the second and third levels require higher attention. Also, depending on the level, certain organizational and technical measures are established to protect personal data (table 2). However, in contrast to the number of checks, such measures require more thorough preparation for pseudonymization procedures at a lower level.

At the first level of threats to the confidentiality of personal data, the problem is solved using typical and standardized methods of information protection. If an attacker does not have access to high-tech hardware and software, then at this level, the guarantee of personal data protection is considered high and reliable.

For the second level of threats to personal data privacy, it is necessary to take into account global models and data flows in it. A statistical analysis of the probability of an undesirable outcome is performed, and the time registration values in the database are deleted or changed. For example, if information about an event known to the attacker was entered into the database, it will not be difficult for the attacker to calculate what information was entered there.

The third level of personal data confidentiality should include the protection of abnormal and rarely encountered data. For example, if an enterprise has an event that occurred only once over a long period of time, an attacker can easily match observations and reveal the identified data of the subject [8,9].

5 Reliability and survivability of the personal data information system

It is also worth noting the possibility of personal data damage due to the failure of the information system. In this case, it is necessary to assess its survivability by determining the probability of an undesirable outcome; if the reasons for failures are divided at the highest (broad) level of consideration.

Thus, the survivability assessment should be considered as the property of preserving all or part of the functions prescribed in the development process with a purposeful destructive effect on the elements and structure of the information system.

The concept of "stability" should be associated with the properties of maintaining the health of an information system in conditions when it operates under external influences that go beyond the tolerance zone allowed by technical documentation. These are mainly physical and chemical impacts, as well as the level of competence and training of the human operator, taken into account in the structure of the functioning information system.

Reliability, as a property to preserve the performance characteristics of an information system within the specified limits when operating in the conditions specified in the technical documentation, is quite accurately defined by GOST And other normative documentation [10-11].

In fact, the state of partial failure implies the possibility of functioning of a faulty information system. At the same time, changes may occur in the information system due to a failure-structural, functional, reliability, and so on. There are risks of the consequences of such changes. These risks are subject to identification, analysis and accounting in the management of the information system and decision-making on strategic further actions. This, in turn, should be based on existing methodological developments for evaluating heuristic States and even intuitive approaches. All this should confirm the need to allocate the mode of operation in the conditions of partial failure, as a specific (independent) direction for evaluating the effectiveness of the tasks defined by the developers when creating the appropriate information system.

In reality, you should point out the following: when designing an information system, developers set tolerances on the parameters of the information system based on the consideration that if the latter have exceeded the established limits, then its performance is preserved [12-13].

At the same time, the operating time is often considered not as the interval between maintenance studies with monitoring and restoration, but as the entire period of planned operation, i.e. before the transition to the limit state. When evaluating decision-making in the period between planned regulations regarding the continuation of operation, the time may be relatively short and the parameters may remain outside the tolerance zone, but not in the critical zone where it is impossible to continue operation [14].

The survivability of an information system is currently associated with its ability to resist targeted destructive influences from individuals who use unauthorized access to it. At the same time, errors in computing and communication processes should also be considered not only as a fact, but also in terms of the amount of damage caused by obtaining an erroneous result. There may be situations when errors can be considered insignificant, and the results obtained remain in the zone of "reasonable solutions"[15].

The concept of stability of an information system can be considered as its ability to perform its assigned functions in conditions similar to those specified in the accompanying technical documentation. For example, this can be attributed to cases when you have to implement a computational process with incorrect input data or interact with an operator who does not have sufficient training in its use.

Conclusion. Said above allows us to focus on issues emerging approaches relating to the protection of personal data of employees by means of pseudonymization, as well as estimates of these properties of the information system as survivability, durability and reliability, especially part of the control devices [16-17].

References

1. K. A. Abdelgawad, K. Yelmarthi, Int. Conf. on Microelectr (ICM), 201–204 (2016)
2. M. Dholu, K. A. Ghodinde, Int. Conf. on Trends in Electr. and Inform. (ICOEI), 339–342 (2018)
3. M. R. Yousefi, A. M. Razdari, Int. J. of Advanc. Biolog. and Biomed. Res., **2(4)**, 473–476 (2014)
4. R. Inglés, P. Perek, M. Orlikowski, A. Napieralski, Mixed Design of Integrated Circuits & Systems (MIXDES), 153–157 (2015)
5. G. Craessaerts, J. De Baerdemaeker, W. Saeys, *Biosystems Engineering*, **106**, 26-36 (2010)
6. D. D. Bochtis, C. G. C. Sørensen, P. Busato, *Biosystems Engineering*, **126**, 69-81 (2014)
7. Z. Zhai, J. Fernán Martínez, V. Beltran, N. Lucas Martínez, Computers and Electronics in Agriculture, **170** (2020)
8. B. Drury, R. Fernandes, M.-F. Moura, A. de Andrade Lopes, Information Processing in Agriculture, **6**, 487-501 (2019)
9. H. El Bilali, M. Sadegh Allahyari, Information Processing in Agriculture, **5**, 456-464 (2018)
10. R. Miodragović, M. Tanasijević, Z. Mileusnić, P. Jovančić, Expert Systems with Applications, **39**, 8940-8946 (2012)
11. D. Bochtis, C. Aage Gron Sorensen, D. Kateris, Operations Management in Agriculture, 79-115 (2019)
12. Ronkainen, *IFAC Proceedings*, **46**, 259-263 (2013)
13. C. Chu, Z. Zuo-xi, K. E. Xin-rong, G. Yun-zhi, IFAC-PapersOnLine, **51**, 346- 352 (2018)
14. L. S. Guo, Q. Zhang, Biosystems Engineering, **91**, 261-269 (2005)
15. M. Kassler, Computers and Electronics in Agriculture, **30**, 237-240 (2001)
16. J. W. Jones, J. M. Antle, B. Basso and others, Agricultural Systems, **155**, 269-288 (2017)
17. S. O'Neill Somers, L. Stapleton, IFAC-PapersOnLine, **48**, 213-218 (2015) 6 *E3S Web of Conferences*, **175**, 05004 (2020) doi: 10.1051/e3sconf/202017505004