

New Technologies for Storing and Transferring Personal Data

Nadezhda Shchegoleva[1,3(✉)], Natalia Zalutskaya[2], Alina Dambaeva[1], Jasur Kiyamov[1], and Aleksandr Dik[1]

[1] Saint Petersburg State University, St. Petersburg, Russia
n.shchegoleva@spbu.ru, {st069509,st080634, st087383}@student.spbu.ru
[2] Federal State Budgetary Institution «Bekhterev National Medical Research Psychiatry and Neurology Center», St. Petersburg, Russia
[3] State Marine Technical University, St. Petersburg, Russia

Abstract. Medical information about a person's state of health, data on the diagnostic evaluation of his symptoms and complaints, especially those related to the field of mental health, require a special reliability of the information storage system, since their disclosure often leads to stigmatization of the patient. The article proposes a new technology for storing and transmitting personal data based on "zeroing the LSB layers" (least significant bit (LSB) component of the Red, Green and Blue color image) and filling these layers with QR-codes containing AES-encrypted personal medical data of the patient. Specified the approach allows you to remove personal data from open access, since the presence of a barcode in the LSB layer practically does not change the image, therefore, without knowing about its presence, an attacker will not pay attention to it, moreover, using existing applications or scanners, it is impossible to read information, and decryption of information is hindered by the lack of an encryption key.

Keywords: Medical data · Barcoding · Color barcodes · Encryption · Data protection · Personal data

1 Introduction

Over the past few years, information technology has been developing at a dizzying pace, and every year the connection between the real and virtual worlds is becoming stronger. The format of communication and information exchange between people and groups of people united by the same interests, government organizations, companies, corporations, etc. has changed significantly. However, unlike the real world, any information that appears in the virtual space is distributed almost instantly, and it is extremely difficult, and sometimes almost impossible, to remove it from the network.

The number of data leaks is growing exponentially every year. News of another company system hack, customer information theft, or demand for payment or ransom is becoming commonplace. The governments of almost all countries are making tremendous efforts to increase the level of information security - a number of new laws are

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13380, pp. 680–692, 2022.
https://doi.org/10.1007/978-3-031-10542-5_47

being developed that provide data protection and provide for liability for violation of the requirements for their processing and illegal distribution [1–5]. Those facts draw increased attention to the methods of storage, processing and distribution of personal information, and also force organizations to strengthen data protection measures in order to prevent their leakage.

According to the Data Leakage & Breach Intelligence (DLBI) intelligence service, medical institutions are among the organizations in which data breaches are most often recorded [6]. Varonis, an information security company, believes that the losses of medical institutions in 2020 from data leaks amounted to $7.13 million, an increase of 10.5% compared to the previous year, and judging by the dynamics of 2021, these losses will increase. This is evidenced, for example, by the appearance in the public domain of data of 300,000 people who recovered from COVID-19, the publication of stolen Pfizer data on the COVID-19 vaccine on the network, etc.

It should be noted that the detection of cyber-attacks on financial institutions is quite fast, but theft and misuse of electronic health records is an almost invisible crime. Since it is usually discovered in an emergency for the owner, when the consequences can be very life-threatening.

An article published on the zdravexpert portal notes that "the life cycle of a leak in the healthcare sector - from the commission of an attack to its detection and remediation - reached a record 329 days in 2020." Obviously, this is due to the catastrophic lack of time to analyze incidents in the health sector, all efforts of which were aimed at combating the pandemic. And in this situation, health care institutions were unable to adequately protect the main thing - the personal data of citizens, including information protected by law on the state of health, examination results and medical recommendations.

Cybersecurity experts note new trends in the black market: information and medical products - various kinds of filing cabinets - are gaining more and more popularity. In January 2022, news broke that the data of 48 million coronavirus vaccination certificates for $100,000 were put up for sale on the darknet. The data of medical policies, medical records, personal data of patients in private and public clinics are put up for sale. Offers are in high demand, from criminal associations for financial gain to pharmaceutical companies, tour operators and funeral service agencies.

It is worth noting a special category of medical personal data related to the field of mental health. Labeling a person as mentally ill often entails serious social consequences, starting with attributing certain negative characteristics to a person, and as a result falling into social isolation and distancing others, and ending with the infringement of rights. Especially when this information falls into the hands of incompetent people who do not have a psychological or psychiatric education, knowledge and experience, but who want to protect other people from a non-existent danger, thus harming an innocent person.

2 Data Encryption

To protect data in practice, several groups of methods are used, including:

– creation of obstacles in the way of the alleged kidnapper, physical and software means.
– data masking or transformation by cryptography and steganography methods.

– development of legal acts and a set of measures aimed at shaping the correct behavior of users.
– creation of such conditions under which the user will be forced to comply with the rules for handling data.

Recently, one of the most effective approaches is data protection in depth, that is, the use of several successive measures to prevent third parties from accessing information. This approach uses multiple levels of data protection, which increases the security of the system and reduces the risk of unauthorized access to information.

In this paper, it is proposed to use two levels of protection. The first is implemented using a cryptographic encryption algorithm that provides data protection by converting it into a form that is incomprehensible and useless for an attacker. The second level of protection is provided with the help of steganography, that is, an algorithm that hides not only the content of the stored information, but also the very fact of storing or transmitting some data embedded in the image. Since a person "at a glance" is not able to accurately estimate the difference between the size of an image without data embedded in it and with data, the main task is to make the embedded data invisible. This approach will be discussed in the next section.

An analysis of the proposed encryption technologies showed that the most promising is the use of symmetric encryption technology. In this case, the block cipher is used to encrypt the identifier using a secret key, which is both a pseudo anonymization secret and a recovery secret.

The advantage of symmetric encryption algorithms is their simplicity, since one key is used for both encryption and decryption, so such algorithms are much faster than their asymmetric encryption counterparts, require less computing power, do not reduce Internet speed, and allow you to quickly encrypt a large the amount of data.

The most promising and one of the most common symmetric encryption algorithms is AES (advanced encryption system) [7], also known as Rijndael. It was developed as an alternative to DES and became the new encryption standard after being approved by NIST in 2001.

AES uses substitution and permutation methods. In the first step, the unencrypted data is converted into blocks, and then encryption is applied using the key. The encryption process consists of various processes such as row shifts, column mixing, and adding keys. Depending on the key length, 10, 12 or 14 such transformations (rounds) are performed. It is worth noting that the last round is different from the previous ones and does not include a mixing sub-process. Thus, AES is faster, more flexible, and more secure because it allows keys of various lengths, and the longer the key, the more difficult it is to crack.

Today, AES is the most popular encryption algorithm - it is used in many applications, including security:

– wireless applications,
– processors and files,
– websites - in SSL/TLS protocol,
– Wi-Fi,
– mobile applications,
– VPN (virtual private network), etc.

3 New Technology for Storing and Transferring Personal Data

To store and analyze the results of various tests and medical examinations of patients at the Research Institute named after V. M. Bekhterev an information system was developed should ensure the protection of personal data, as well as research results containing information that allows you to identify the patient directly or indirectly. This is an extremely difficult task, since in future this system should be integrated with other brain research systems within the framework of international collaborations.

Currently, bar codes are widely used in most medical institutions for fast and accurate management of patient flows. In information systems, barcodes are used to register and record patients, generate medical records and referrals for procedures, extracts and prescriptions, while there is no explicit information about patients on the documents.

An analysis of the characteristics of existing barcodes shows that QR-codes (Quick Response) are of the greatest interest [8, 9], having the highest capacity for recording information from ASCII, Cyrillic characters and, moreover, allow recovering up to 30% of lost information. The generation and reading of this kind of barcodes is quite simple, there are many different kinds of sensors and applications on the market that allow reading using mobile devices. Of particular note the recently appeared color QR-codes, which actually consist of 3 QR-codes, so their capacity for recording information is three times higher. An additional advantage is the description of QR-code as ISO standards, which does not require any royalties.

However, when realizing access to information, financial and material resources within the framework of a corporate network through barcodes in a significant number of medical institutions, there is an almost open transmission of patient information through communication channels. In this case, there are many opportunities for intercepting confidential and personal data, their substitution or replacement of these data by modifying them, since the barcode decryption algorithms are open and the programs implementing this process are also in the public domain.

Therefore, in order to protect the personal data of patients in accordance with the adopted laws, an additional step of hiding data from being read by illegal users should be implemented.

In [10] presents a method and algorithms for embedding QR-codes with embedded biometric and documentary information in color images to form BIO QR-codes.

The idea of the algorithm proposed by the authors is based on “zeroing the LSB layers” (least significant bit (LSB) of the Red, Green and Blue components of the original color image and filling these layers with the corresponding QR-codes. This is what makes it possible to remove personal data from public access, since using existing applications or scanners is impossible on the one hand, and on the other hand, the presence of a barcode in the LSB layer practically does not change the image, therefore, without knowing about its presence, an attacker will not pay attention to it.

4 Algorithm for Embedding Personal Information in a Color Image

Considering the recommendations [11], the process of embedding personal information in a color image will consist of the following steps.

1. Preparation of messages for QR-codes with documentary information.
   To form a message that will be used to generate a QR-code, it is necessary to group the initial personal textual and digital information (for example, full name, year and place of birth, profession, etc.), biometric data information (weight, height, etc.), examination results, established diagnoses, etc. by arranging them into semantic groups to record them in three barcodes. At the same time, separator characters (for example, #) can be used to structure information.
   Such an organization of the message makes it possible to preserve the semantic structure of biometric and documentary information. This makes it possible to simplify the process of generating QR-codes from them and decoding them to read information to the recipient. Depending on the purpose of using the information, it is possible to generate QR-codes containing different sets of personal data and documentary information, while embedding one, two or three codes in the corresponding LSB layers of the Red, Green and Blue components.
   In the original algorithm, the next step is to write the generated messages into QR-codes, but in this case, the information will be stored in its original form and can be easily read. Therefore, as shown above, in order to protect information, it must be encrypted. Therefore, a modification of the algorithm is required.
2. Encryption of messages for QR-codes with documentary information
   The results of the study showed that the most efficient, secure, fast and flexible encryption algorithm is AES, the advantage of which is the ability to use keys with different lengths, which can be selected based on user needs and system limitations.
   Therefore, the next step is to encrypt prepared messages for QR-codes with documentary information using the AES algorithm and generate a key, the length of which is selected by the user in accordance with security requirements, and contains the parameters of the QR-code in the image.
3. Writing encrypted messages to QR-codes
   Graphically, QR-codes are halftone images in ".png" format. In [10] it was shown that this image format provides a representation without compression noise, full preservation of information in memory exchange operations, which means that the QR-code will not be distorted by the noise of restoring the recorded information.
4. Preparing the container image
   The container can be any color image, the size of which exceeds the size of the barcode. To do this, the input color image is decomposed into three components R, G and B, which are three matrices of size $M \times N$ (where M and N are the size of the image in pixels). In graphical representation, we get three halftone images - components R, G and B, which are shown in Fig. 1.

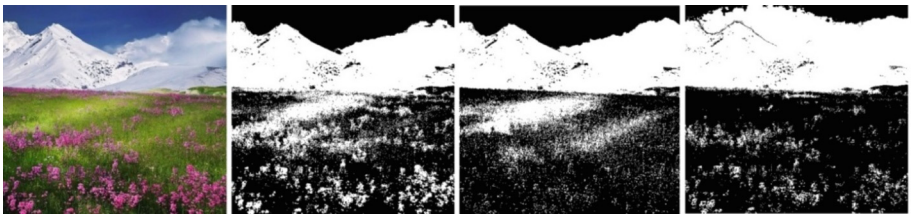


Fig. 1. Original color image and the result of its decomposition into components R, G, B.

From halftone images, the R, G and B components are removed in the LSB (least significant) layer, the part whose size corresponds to the size of the embedded QR-code.

5. Formation of the container image.

In place of the removed part of the LSB layer of halftone images, the R, G, B components and the contents of the QR-codes are embedded. Next, the reverse operation is performed - combining all the layers of grayscale images and the R, G and B components into one color image that contains personal information in the form of QR-codes. The result of the process of embedding QR-codes in a color image is shown in Fig. 2.

Figure 3 shows a diagram of the process of embedding information in color images and the stages of information transformation: defining a set of personal data, preparing messages for each component R, G, B, AES encryption, generating QR-codes, embedding the corresponding bar codes in the components.

Reading information from a filled container image is carried out in reverse order (from right to left) by extracting embedded QR-codes from the LSB layers of components, decoding QR-code messages, decrypting messages, reading message information in accordance with the accepted record structure.

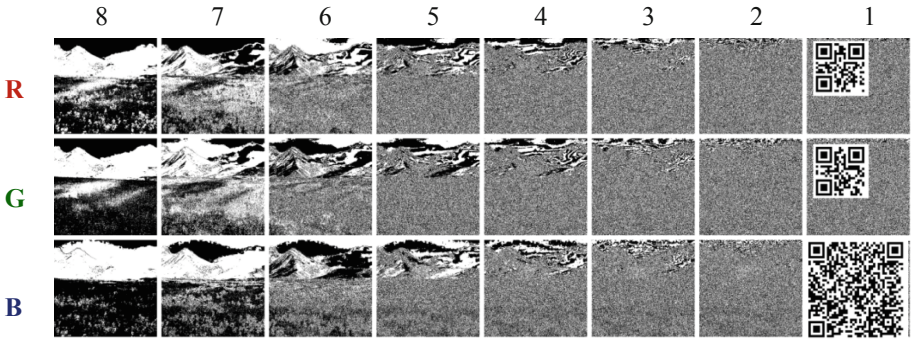


Fig. 2. R, G, B components of the "filled container image".

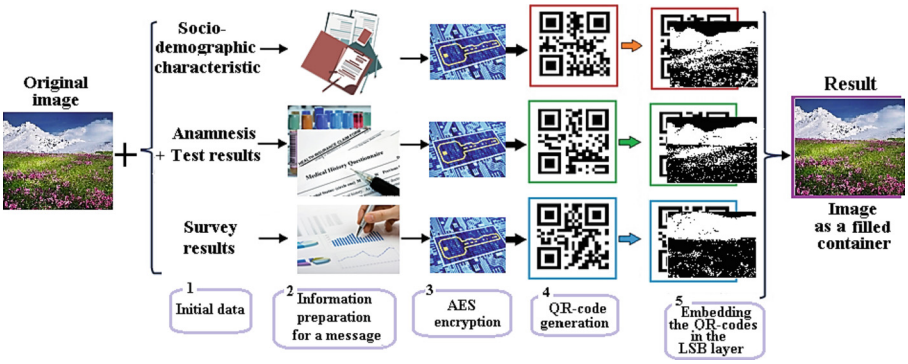


Fig. 3. Scheme of the process of embedding information in color images.

5 Selecting a Container Image for Storing and Transferring Data

The purpose of forming a container image is to hide information during the transfer and storage of personal data. However, as experiments show, not every image can be used for these purposes. This is due to the fact that all barcodes recorded in one area of the LSB layers can lead to a slight distortion of the original image. At the same time, this distortion may not be noticeable to an ordinary user, and at the same time, a specialist will notice a difference, which can be explained by errors in data transmission, but re-sending will not correct the situation, which will lead to increased attention to this image, and this contradicts the task. Example of such image is shown in Fig. 4a. This problem can be solved in different ways, which will be discussed in this section.

Image Selection Based on Color Histogram. The first and easiest way is to choose an image that will change slightly when barcodes are added to the LSB layers. To do this, we calculate the color histograms of the image obtained from the original color image. In this case, the larger the range of values, the more suitable the image will be for using it as a container. Figure 4 shows sample images and their color histograms. At the same time, Fig. 4a shows an example of an image in which the appearance of a barcode will be noticeable, Fig. 4b shows an image in which distortions will be hardly noticeable, and Fig. 4c shows an image that should be chosen as a container.

Selection of the Content of the Barcode. Let's consider the content of the barcode in more detail. A QR-code consists of six components: indentation, search pattern, alignment pattern, sync bars, version, data cells. The first two components - the indent and the search pattern (see Fig. 6, highlighted in color), which correspond to the largest areas that have one color, which can distort the original image, have the greatest importance for the formation of the barcode image.

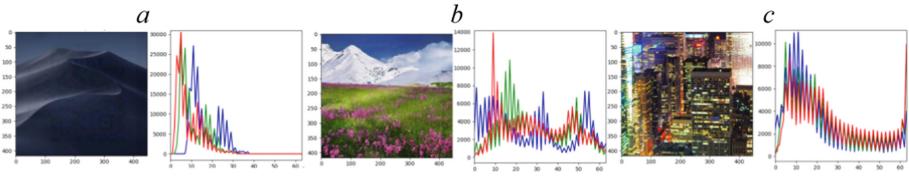


Fig. 4. Color histograms of images.

The padding is the white border around the outer edge of the QR-code. Without this frame, the scanner will not be able to determine what needs to be read - the surrounding elements will interfere with it.

The search pattern is three black squares: in the lower left, upper left and upper right corners, they show where the outer borders of the QR-code are and how it is oriented.

These components have a size strictly limited by the standard, so they can be removed before being placed in the container and added when reading the barcode. The content part, namely the data itself, will then have the form shown in Fig. 5 right. Obviously, such a barcode image is already becoming less noticeable in the original image.



Fig. 5. Separation of barcodes and options for their recording in the image-container.

It should be noted that in this case, the information can also be read using the modification of the content of the barcode described above. Therefore, the next step, which will make the code invisible, will be dividing it into several parts and writing it to different parts of the image, and the positions for writing in each layer may be different. An examples of possible ways to separate barcodes (see Fig. 2) and options for writing them to the container image is shown in Fig. 6.

Barcode Image Conversion. Another approach that allows you to hide the information contained in a barcode is the use of geometric transformations of image matrices: rotation by an angle that is a multiple of 90°, matrix transposition, rotation around an axis that coincides with one of the rows or columns, etc. Several examples of such transformations are shown in Fig. 7.

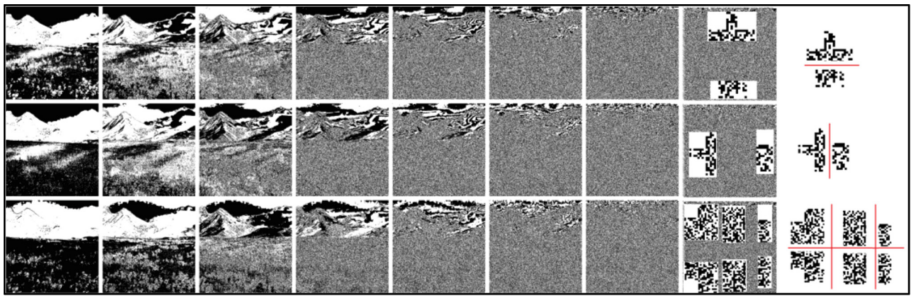


Fig. 6. Examples of barcode image transformations.

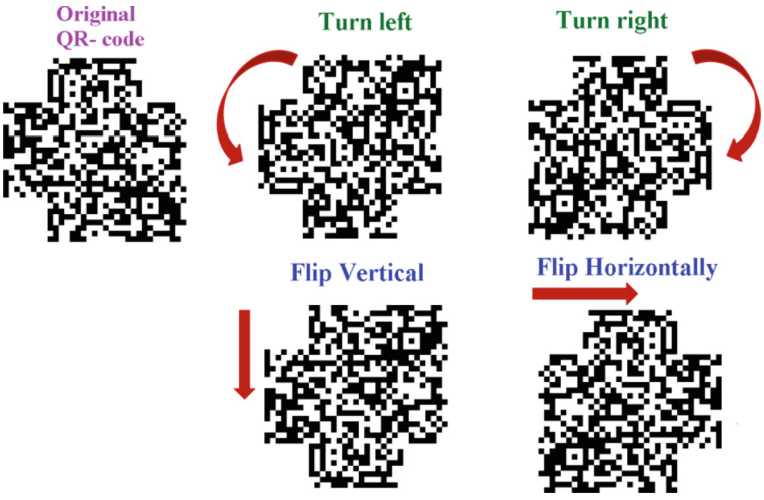


Fig. 7. Examples of barcode image transformations.

Similarly to the transformation of the barcode image, it is possible to perform transformations of its parts shown in Fig. 6. In this case, for each part, its own method of change can be used.

Note that the image shown in the pictures is small (446 * 417 pixels), so you may notice a slight distortion of the picture that looks like “broken” pixels. However, if the image has a significantly larger size, then the human eye will hardly see the changes made. And in this case, such a change will serve to hide, and therefore protect the data placed in the container image.

At the same time, the proposed methods for changing the barcode image do not require complex mathematical transformations, graphic libraries and have low computational complexity, which allows them to be used for a wide range of data transmission and storage systems.

Until now, the question has remained aside, how does the recipient of such an image-container with the patient’s personal data find out about the performed barcode conversions?

The solution to this problem can be implemented either by generating an encryption key in such a way that it contains values corresponding to certain transformations in certain places, and then the recipient of the message can perform all the necessary actions. Other way around is on the basis of a preliminary agreement between the sender and recipient of the order and list of transformations to be performed. The combination of the above approaches is also possible.

In view of the foregoing, the proposed new technology for storing and transferring personal data will consist of the following steps (see Fig. 8):

1) Preparation of messages for QR-codes with documentary information.
2) Encryption of messages for QR-codes with documentary information
3) Selection of the content of the barcode image
4) Dividing the selected content part of the barcode image into several parts
5) Application of geometric transformations of image matrices of barcode parts (rotation by an angle multiple of 90°, matrix transposition, rotation around an axis that coincides with one of the rows or columns, etc.)
6) Writing the converted parts of the barcode to different areas of the container image.

Accordingly, the extraction of information is carried out in the reverse order.

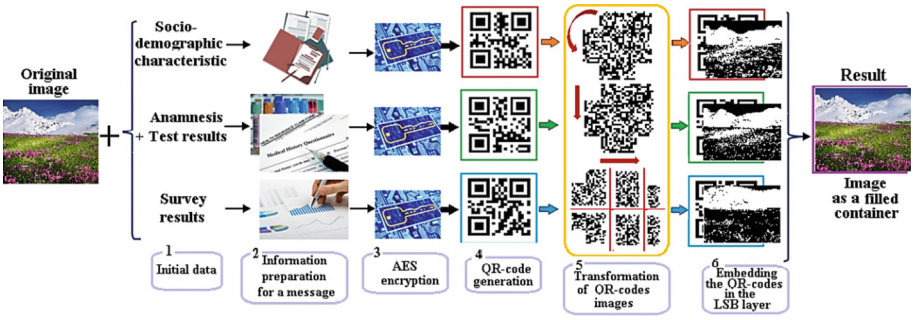


Fig. 8. New technology for storing and transferring personal data.

6 Practical Implementation

The new data storage and transmission technology proposed in this article was implemented as an application consisting of two modules - for the sender and the recipient. An example of the application is shown in Fig. 9 - data input for forming a container image and Fig. 10 - the process of extracting data from a container image.

At the same time, as can be seen from the presented figures, it is enough for the sender of personal data to simply select the necessary information from the Excel table, specify the encryption key and select an image to form the container. Knowing the key, the recipient can easily extract information from the container image. It is worth noting that the original image has not changed much and without knowing about the presence of any information in it, the attacker will not pay attention to it, and therefore we will be able to safely store and transfer personal medical data.

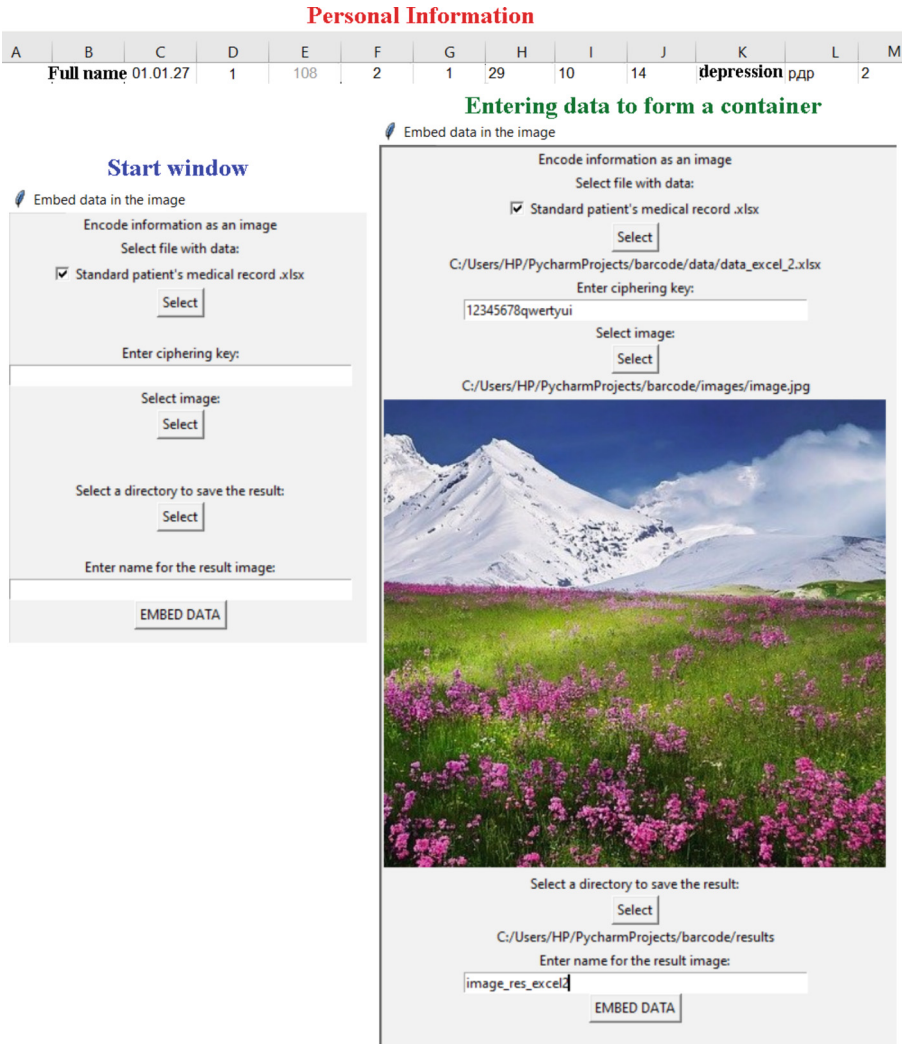
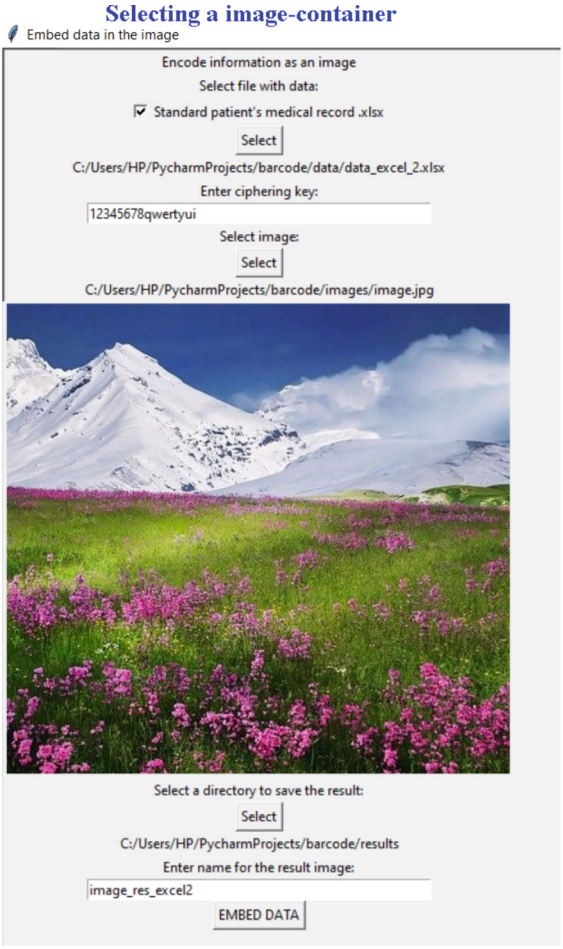


Fig. 9. Data input for the formation of an image-container.



The result of extracting data from the image-container

A	B	C	D	E	F	G	H	I	J	K	L	M
	Full name	01.01.27	1	108	2	1	29	10	14	**depression**	рдр	2

Fig. 10. The process of extracting data from a image-container.

7 Conclusion

It should be noted that container images obtained in this way do not make it possible to detect the existence of documentary information in them. The use of color images with QR-codes embedded in them containing encrypted information as a storage and carrier of personal data will ensure the protection of this information, as well as restrict open access to it, since the embedded information can be read and decrypted only if the encryption key is present and knowledge of the structure of information is known.

The proposed new technology will make it possible to apply a universal standard when transferring personal data over a network, exchanging information, including through international channels, without losing any part of it and/or replacing it, as well as significantly reduce the losses of medical institutions from data leaks and significantly increase security of personal data processing.

Acknowledgments. The research is partially funded by the Ministry of Science and Higher Education of the Russian Federation as part of World-class Research Center program: Advanced Digital Technologies (contract No. 075-15-2020-903 dated 16.11.2020).

References

1. General Data Protection Regulation – GDPR
2. NIST.SP.800-122. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)
3. NIST.SP.800-53. Security and Privacy Controls for Federal Information Systems and Organizations
4. NIST.SP.800-37. Risk Management Framework for Information Systems and Organizations A System Life Cycle Approach for Security and PrivacyNIST.IR.8053. De-Identification of Personal Information
5. NIST PRIVACY FRAMEWORK 1.0 (PRAM): A tool for improving privacy through enterprise risk management
6. Data Leakage & Breach Intelligence. https://dlbi.ru/data-base-leaks-risks/
7. Advanced Encryption Standard (AES)
8. Kukharev, G.A., Kaziyeva, N., Tsymbal, D.A.: Barcoding technologies for facial biometrics: current status and new solutions. Sci. Tech. J. Inf. Technol. Mech. Opt. 18(1), 72–86 (2018). https://doi.org/10.17586/2226-1494-2018-18-1-72-86. (in Russian)
9. Kaziyeva, N., Kukharev, G., Matveev, Y.: Barcoding in biometrics and its development. In: Chmielewski, L.J., Kozera, R., Orłowski, A., Wojciechowski, K., Bruckstein, A.M., Petkov, N. (eds.) ICCVG 2018. LNCS, vol. 11114, pp. 464–471. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00692-1_40
10. Nazym, K.: Methods and algorithms of barcoding for facial biometrics. Academic Dissertation Candidate of Engineering (2020)
11. Patent of the Russian Federation No. 2714741. Method for generating a color QR-code according to images of faces and a device for its implementation. In: Kukharev, G.A., Kazieva, N., Shchegoleva, N.L.