university of groningen

University of Groningen
Faculty of Science and Engineering

Bachelor Thesis in Computing Science

Is pseudonymisation the silver bullet for compliance with the GDPR?

Author:

Sander de Jong
s2013738

Supervisors:

dr. F.B. Brokken
prof. dr. G.R. Renardel de Lavalette

Advisors:

mr. E. Hoorn
ing. V.A. Boxelaar
R.J. Oostergo

August 6, 2017

Academic Year 2016/2017

Abstract

The GDPR imposed by the European Union will be active from May 2018 and will place more requirements on data collected for various purposes. Pseudonymisation can help protect these data. Pseudonymisation techniques include encryption, hashing and tokenisation. While tokenisation provides the best security by relying on non-mathematical principles and minimising access to the central token storage, encryption is very useful for sharing pseudonymised data. While transferring data, a private key is sent to the receiver to decrypt the data. Hashing is less secure but an adequate technique when resources are limited.

Even when data is pseudonymised, it can be traced back to individuals. Therefore, consent has to be gained from selected data subjects. Even when consent has been offered, data collected should be minimilised adhering to 'data protection by design'. This principle states that in every step of data collection and processing, data minimilisation should be taken into account. Providing training to (long-term) employees will encourage the use of pseudonymisation techniques and data minimisation.

Contents

Chapter 1

Acknowledgments

We would like to thank Frank Brokken and Gerard Renardel de Lavalette for their supervising role during this project. Especially Frank for the weekly meetings and the detailed feedback he provided to our earlier drafts.

We would also like to thank Esther Hoorn for the initialisation of this project as a whole. We have enjoyed the collaboration between different faculties of the University. Another appreciation is for Vincent Boxelaar for taking the time to meet with us, explain current security problems and describe the university project. Lastly, we would like to thank René Oostergo at the UMCG for the meeting in which we exchanged and discussed security issues in the different fields.

All chapters with prefix 'project' are written by Erik Bijl, Sander de Jong and Victor Preda as part of the overarching project "Towards a catalogue of data protection measures for research projects". All other chapters are written solely by Sander de Jong.

Chapter 2

Project Introduction

This report is the result of a project of computing science students in which the authors analyse the General Data Protection Regulations (GDPR) from a technical perspective. In this report the GDPR is analysed and technical implementations are extracted, which are covered in individual theses. The three technical implementations taken from the GDPR are the following: encryption, anonymisation and pseudonymisation each of which is covered by one of the authors. In their theses the authors look at the relation between the GDPR and their technique, analyze how this technique can be used regarding the GDPR, and analyze certain methods to accomplish this technique.

The research starts with 2.1 and explains what the General Data Protection Regulations (GDPR) [3] is and what its consequences are. As an answer to the GDPR, a model to protect data was published called the Standard Data Protection Model (SDM) [4]. The SDM builds a bridge between the regulations described in the GDPR and some generic implementations. This is done by establishing protection goals from the GDPR and specifying measures to ensure these protection goals. Section 2.2 describes those protection goals and measures after which they are discussed to see whether those measures are useful for the purpose of our theses: finding technical implementations of the regulations described in the GDPR. After the bridge towards technical implementations is built, chapter 3 extracts three technical techniques that help to comply with the GDPR. These techniques are: encryption, anonymisation and pseudonymisation. The three techniques are introduced and analysed by each individual author in the later chapters. The findings of every individual subject is discussed in chapter 8 and concluded with a case study on a existing research project.

2.1 General Data Protection Regulation

"Personal data is the new oil of the Internet and the new currency of the digital world." was said by the European Consumer Commissioner Meglena Kuneva in March 2009. It turned out to be true, in 2013 the Financial Times [1] published a tool to calculate how much your personal data is worth. General information about a person such as age or gender was worth $0.0005 per person. In Europe, regulations to protect personal data were adopted by the EU in 1995 and collected in the Data Protection Directive [2]. The Directive is no longer sufficient to deal with digital trends and technologies that emerged in the last decade. Therefore the EU has adopted The General Data Protection Regulation [3] to ensure the necessary data protection. The General Data Protection Regulation (GDPR) was approved by the European Parliament and the Council in May 2016. After a transition period of two years the GDPR will be active starting 25 May 2018. By means of the GDPR, the EU Parliament and Council protects their citizens with regard to the processing of personal data and the free movement of such data. Although the key principles of data privacy still hold true to the Directive of 1995, many changes have been proposed to the regulatory policies. In the GDPR individuals from which data is collected are defined as data subjects. What data is collected and the purposes for these data are defined by data controllers. The term data controller "means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data". The data controllers assign data processors. The term data processor "means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller"

One of the obligations that is stated in the GDPR, is the introduction of a data protection officer. Every organisation that collects personal data must designate a data protection officer who makes sure that the GDPR is applied. This data protection officer has, according to Article 39 of the Regulation [6], the following tasks:

1. Informing and advising the controller or processor and its employees of their obligations to comply with the GDPR and other data protection laws.
2. Monitoring compliance with the GDPR and other data protection laws, including managing internal data protection activities, training data processing staff, and conducting internal audits.
3. Advising with regard to data protection impact assessments when required under Article 35 [6].
4. Working and cooperating with the controller's or processor's designated supervisory authority and serving as the contact point for the supervisory authority on issues relating to the processing of personal data.
5. Being available for inquiries from data subjects on issues relating to data protection practices, withdrawal of consent, the right to be forgotten, and related rights.

When organisations do not comply with the GDPR high penalties are applied. Therefore a global trend emerged to comply with the GDPR. Organisations are monitoring and revising their data-flow. From 2014, the year of acceptance by the EU, manuals are published to comply with the GDPR. In these years an increase in available research emerges due to the fact that organisations have to change their current data management. Research data will fall under these new regulations and therefore universities must prepare for the new requirements of the law. As a result the University of Groningen started a project with law and computing science students to analyze the GDPR. This paper is the result of the Computing Science project and was conducted from a computing science point of view. This research helps researchers and data protection officers by providing a technical analysis and conclusion regarding the technical aspects in order to comply with the GDPR.

2.2 Standard Data Protection Model

As an answer to the new regulations a model called the Standard Data Protection Model (SDM) was published on the website of the German "Datenschutz Centrum" [4]. The SDM converts the legal requirements into a set of technical and organisational data protection measures. In this chapter the requirements set by the regulation are evaluated and criticised.

The SDM describes the general protection goals that need to be satisfied in order to adhere to the regulation. These protection goals consist of the overarching research goal of data minimisation followed by the three classical protection goals in data security (availability, integrity and confidentiality) and three protection goals aimed towards the protection of data subjects [5]. The SDM does not yet give a concrete technical implementation to satisfy the protection goals. This part of the SDM is still forthcoming but the date of its addition is yet to be specified. Although the SDM describes some generic measures that can be used to fulfill the protection goals, these measures are directly translated from the legal document and not all of them are useful in practice. In this section these measures are mentioned and analysed from a technical perspective.

2.2.1 Data minimisation

The overarching protection goal data minimisation states that data collection should be limited to the essential. During the research process as a whole, as well as every step involved in it, measures need to be taken to make sure that just the data which is strictly necessary for the research is collected and processed. This is defined in particular for every research project. Even before taking procedural and technical measures, controllers should evaluate whether collecting the proposed amount of data is really necessary. Additionally, controllers should limit the number of

parties having access to the data and the control they have over the data. This goal should be key throughout the entire organisation and the system should be built around it. It should limit data usage during its entire lifecycle, from collection to processing ending with deletion or complete anonymisation.
The following are some generic measures proposed by the SDM:

1. Reduction of gathered data from subjects and options to process the data.
2. Limit options to access the data.
3. Preferably use automated processes to limit the possibility of interference by not using processed data.
4. Implement options to block or erase data.
5. Implement pseudonymisation and anonymisation.
6. Implement options to change the procedures for processing data.

Controllers should limit the amount of data which is collected, the amount of processors that have access to the data and the control these processors have over the data. The amount of data can be limited by omitting certain data fields or attributes. Data can be further minimalised by erasing the data as soon as possible or transforming them using anonymisation or pseudonymisation.

2.2.2 Availability

The first of the traditional protection goals is called Availability. This is the requirement to have data accessible, comprehensible and processable in a timely fashion for authorised entities. The data must be available and can be used properly in the intended process. An authorised user must be able to find, access and interpret the data. Therefore even if a user could find the data, but has no possibility to interpret the data, this rule is violated.
The following are some generic measures proposed by the SDM:

1. Preparation of data backups, process states, configurations, data structures, transaction histories etc., according to a tested concept.
2. Protection against external influences (malware, sabotage, force majeure),
3. Documentation of data syntax.
4. Redundancy of hard- and software as well as infrastructure,
5. Implementation of repair strategies and alternative processes.
6. Rules of substitution for absent employees.

This measures may look straightforward and don't have a technical perspective. On the other hand three measures look important to us. Redundancy helps to increase the reliability of hardware and software. Data backups are highly relevant because they will ensure that certain states of the data will be stored safely. Raw data without explanation is hard to interpret so documentation of data syntax should be provided in order to help interpreting the data correctly and will contribute to the availability.

2.2.3 Integrity

The second protection goal refers to both data processes and systems and the actual data itself. Information technology processes and systems must at all times comply with the specifications that were established for the execution of their intended use. On the other hand the data must be up-to-date, authentic and complete. Integrity means that the data must be unmodified, authentic and correct.

The following are some generic measures proposed by the SDM:

1. Restriction of writing and modification permissions.

2. Use of checksums, electronic seals and signatures in data processing in accordance with a cryptographic concept.

3. Documented assignment of rights and roles.

4. Specification of the nominal process behaviour and regular testing for the determination and documentation of functionality, of risks as well as safety gaps and the side effects of processes.

5. Specification of the nominal behaviour of workflow or processes and regular testing of the detectability respective determination of the current state of processes.

Integrity measures are mostly technical and play a role in our research. As stated in the previous chapter integrity means that the data must be unmodified, authentic and correct. Checksums will show differences when the data is modified and therefore guarantee that unknown modifications of data are discovered and can be dealt with. Signatures guarantee that the data is not modified, or at least shows who modified it, therefore it ensures authentication and exposes unauthorised modifications.

2.2.4 Confidentiality

Confidentiality means the need for secrecy by limiting the number of parties who have access to the data and the non-disclosure of these parties. To ensure the confidentiality of a research project, only parties which are authorised should have access to the data. This is not only violated when a third party, unknown to the controller, gains access to the data, but also when a party known to the controller has acquired access for the wrong reasons. Taking into account 'privacy by design', the controller should only give access to parties which are related to the research project and inevitably need to have access to process the data and are authorised.

The following are some generic measures proposed by the SDM:

1. The controller defines the rights and role of the processors according to the principle of necessity. Also define the procedures, regulations and obligations.

2. Implement a secure authentication system.

3. Specify the use of available resources by the data processors.

4. Protect the data against unauthorised access by implementing encryption and protection against hacking.

5. Specified environments (buildings, rooms) equipped for the procedure,

6. Specification and control of organisational procedures, internal regulations and contractual obligations (obligation to data secrecy, confidentiality agreements, etc.).

Most of these measures are fairly straightforward. Organisational measures and preparation of the working environment for secure data processing are outside the scope of this paper. Of the more technical measures, the importance of encryption and limiting the available resources is highly relevant because it is mentioned for other protection goals as well. The need for a secure authentication system and protection against hacking is important but straightforward for computing scientists.

2.2.5 Unlinkability

Unlinkability means that data should only be processed and analysed for the purpose for which it is collected. It ensures that data is not linked across different domains and research projects. The following reasons are given to support the idea of allowing linkability:

1. Archival purposes that are of public interest.

2. Scientific or historical research purposes.

3. Statistical purposes.

In all these cases safeguards have to be in place to ensure the rights and freedoms of data subjects. Data minimisation and pseudonimysation are examples of these protective measures.

The following are some generic measures proposed by the SDM:

1. Restrict the processing of data and transfer rights.
2. In terms of programming, omitting or closing of interfaces in procedures and components of procedures.
3. The controller defines clear roles and gives people access to the data accordingly.
4. Define procedures processors can follow within interfaces to make sure that processors know what they can and can't do.
5. Clearly define the boundaries between departments and organisations.
6. Approval of user-controlled identity management by the data processor
7. To make sure that data cannot be linked back to a data subject while linking databases, use purpose specific pseudonyms, anonymisation services, anonymous credentials, processing of pseudonymous or anonymous data.
8. Regulated procedures for purpose amendments.

Again the importance of a clear organisation under the supervision of a data protection officer is mentioned and the restriction of access to the data. Since unlinkability enables transfers between different research groups, the measures include limiting access to these transfer operations. Only authorised personnel should be able to transfer data. To enable these transfers, the data has to be pseudonymised or anonymised. Because these are purely technical measures and they are key in achieving multiple protection goals they will be the main focus of this paper.

2.2.6 Transparency

The Transparency goal has as its main purpose tracking the data which is collected and all the details which concern it. In order for this goal to be met some details have to be brought to the attention of the research subjects. The subjects must be aware exactly of the data which is collected, the purpose for which the data is collected and the parties to which this data could potentially be disclosed and the processes which are performed on this data. These are the main details which need to be properly documented and easily be available for the subjects and the controller.

The following are some generic measures proposed by the SDM:

1. Good documentation for all the details concerning the research (consents, objections, data flows, IT systems used, operating procedures etc).
2. Verification of the authenticity of the data sources.
3. Keeping logs of access and modifications.

For this protection goal the first measure proposed is good documentation. This fact can be easily overlooked but as many programmers and software engineers have felt if the documentation is of poor quality it takes much more time and resources to achieve the desired goal. In this case the documentation could be considered also a gateway in the particular project about which it is written. If good documentation is available to someone who wants to get information from the project or use it in their own project etc, it is much easier for that person to figure out how to interact with the information in the project. Naturally what comes easiest to a person is what is familiar, therefore the documentation could be formated in a standard format. This format could be designed and then used as a template in a program which would be available to anyone who would need it in order to document all the details about the project that they work on. For the second measure from a technical point of view a certification-like system could be used. In the human society a social contract is present: if a website has a valid certificate issued from a recognized institution then the website can be considered as secure. A similar solution could be proposed in this case as well. Lastly the third measure could be approached in the same way that a server keeps logs for its files and rights to them. It could be a straight forward implementation in a UNIX-like file permission system.

2.2.7 Intervenability

The Intervenability goal has as its main purpose to ensure the ability of the subject to review the data collected about him/her and be able to correct, restrict access and/or erase any part of it. This ability should, ideally, be provided in an easy and quick fashion by the controller. Furthermore this ability of the subject can be restricted under specific legal cases which are mentioned in the SDM and the GDPR but this fact lives outside the technical purpose of the paper. The technical aspect of this goal can be better achieved with the help of realising at a high standard the technical implementation of the above-mentioned goal, Transparency. If the data has been shared with any third party the controller is required to guarantee that any correction, restriction or deletion of the data by the subject is propagated to all the points at which the data was replicated in a timely fashion.
The following reasons are given to support the idea of allowing intervenability:

1. Establishing a Single Point of Contact (SPoC) for data subjects.
2. The technical ability to compile modify or erase completely data about any one person
3. The ability for the controller to keep track of all the data.
4. Having a module-like system in which individual functionalities can be disabled without affecting the whole system.
5. Documentation about the security system and the data protection measures.
6. Documentation about handling of malfunctions, changing of procedures and problem-solving.

The measures proposed above span across multiple concepts and therefore a much more in-depth research is needed in order to offer a proper and documented technical solution, because of this these measures will not be explored in great depth in this paper. However, some starting ideas are given below in order to offer a starting point for any interested parties. The first three measures could be implemented in the same fashion as the Andrew File System. Two processes can be used to implement the system, in the case of the file system these processes are named Venus and Vice, which can keep track and manage all the files which are requested from the server. The server would be the SPoC in the technical sense and its administrators would be the SPoC on the human side. The fourth measure could be implemented alongside the lines of how API's are implemented in the current market. The last two measures do not fall strictly in the territory of computing science but are a general requirement for any good system. There are several philosophies on writing and maintaining good documentation about a system.

Chapter 3

Project Methods

In the previous chapter we introduced the General Data Protection Regulation (GDPR) and the model developed to meet this new regulation called the Standard Data Protection Model (SDM). The SDM is based on the protection goals we introduced. In order to develop a system, unfortunately, it is impossible to fully satisfy all these protection goals. This is a result of the tension between these protection goals. As an example take confidentiality versus availability. On the one hand a system having a high degree of confidentiality could mean that there are different ways of authenticating users. On the other hand in order to have a high degree of availability, the data has to be accessible in an easy and straight-forward fashion. One can easily observe that having a high degree of confidentiality results in having less availability. If a system is fully transparent it could mean that it is clear how the data is processed but this will make it easy to link data to real subjects. Therefore personal data can not be truly unlinkable. We conclude here that when building a system, choices must be made to satisfy the protection goals and, at the same time, balance the importance of each goal with respect to the situation at hand. Arguments should be given to explain which, why and how you satisfy the protection goals.

We have researched the protection goals and their measures according to the GDPR. In the following chapter we will present some technical implementations in accordance with the suggestions in the GDPR.

We consider that encryption, anonymisation and pseudonymisation are three security concepts which are fundamental to a system that processes personal data and they also help in achieving an acceptable level for most of the protection goals which are important motives for both the GDPR and the SDM. Because of this these three concepts are presented in more detail in the coming chapters. We should also mention that in both of the above-mentioned official documents, encryption, anonymisation and pseudonymisation are introduced and mentioned but no technical details are specified. This paper will offer support in establishing how these abstract notions should be implemented in real cases and will provide at least entry-level information about the facets of these three cornerstones of security.

3.1 Encryption

Encryption is the process through which data are transformed into unintelligible text which to the human eye seems like random characters. This is done to avoid that an unauthorised third party can understand the data. Decryption is the reverse of the process of encryption. When dealing predominantly with user data one could easily arrive at the conclusion that encryption is an important part for the protection of privacy of each subject. The same conclusion was reached also by the European Parliament, therefore the inclusion of encryption as a measure for privacy can be observed in Article 6 paragraph 4(e) of the GDPR. Little attention is payed to this first appearance and one can think that encryption only is named as an example. These thoughts should be muted by the next appearance of this concept in Section 2 entitled "Security of personal data". Article 32 lists the concept presented above and encryption as technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate.

3.2 Anonymisation

Anonymisation is a technique to anonymise data in a particular manner such that no person can be identified from these data. As a result the data is not be seen as personal data and will be placed outside the scope of the GDPR. This is regulated in recital 26:

The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.

When data is anonymised, there is no appearance of personal data and the GDPR will not play a role. This could be really helpful for researches to handle their data. Anonymiation is done correctly when one is unable to make a profile of a subject from said subjects data. According to the GDPR article 4(4) profiling is defined as:

'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.

3.3 Pseudonymisation

Pseudonymisation is the reversible technique of anonymisation. According to the GDPR article 4(5) the definition of pseudonomisation is:

pseudonymisation means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person

Pseudonomisation is a technique to replace identifiable fields (such as names, phone numbers, email addresses etc.) with pseudonyms in order to separate the identifiable fields from the actual data. For the use of pseudonomisation we look at recital 28 of the GDPR. This states:

the application of pseudonymisation to personal data can reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations.

Moreover according to the GDPR pseudonymisation is seen as a measure of data protection. In Recital 78 it is stated that

The controller should implement measures which meet in particular the principles of data protection by design and data protection by default. Such measures could consist, inter alia, of minimising the processing of personal data, pseudonymising personal data as soon as possible.

In article 32(1) pseudonomisation is seen as an appropriate technical and organisational measure to ensure a high level of security. Article 89(1) says that pseudonomisation may be included in measures to ensure the principle of data minimisation. We can conclude by saying that according to the GDPR, pseudonymisation is a new and highly useful measure to ensure data protection. Although recital 28 states:

The explicit introduction of 'pseudonymisation' in this Regulation is not intended to preclude any other measures of data protection.

Chapter 4

Introduction

In 2018, the EU activates its General Data Protection Regulation (GDPR) [3]. It will replace the existing legislation concerning data protection, the Data Protection Directive [9] (DPD) introduced in 1995. Since the GDPR will only be mandatory from May 2018 onwards, security advisers are still figuring out the best way to implement the measures in their data systems. While there are many articles and blogs available that analyse the regulations from a legal standpoint, articles that cover the technical implementation are missing.

The German Standard-Datenschutzmodell [4], covered in the introductory document *"Towards a catalogue of data protection measures for research projects"*, is an example of a legal analysis of the GDPR and has complemented this analysis with a model that can be used to comply with the regulations. This model has been analysed in the introductory chapter and the most important technical implementations have been extracted from the regulations. These measures can be divided into two categories: De-identification and encryption. A combination of these two categories of measures has to be used to comply with the GDPR.

4.1 Definitions

A data subject is a natural person participating in research whose personal data is processed by a controller or processor [10]. Every project should have a designated data officer who makes sure that the research is done in compliance with the GDPR. Data controllers are appointed by the data officer to determine the purposes, conditions and means of the processing of personal data [10]. A processor is a person assigned by a data controller to handle the collection, recording and use of personal data. An identifier is an attribute in the data records that could link back to an individual data subject's name or address. Attributes are also considered identifiers when they do not directly link back to an individual but can be used to identify a data subject while combined with other attributes from the same dataset or other datasets [11]. To ensure the privacy of the data subjects every identifier has to be removed from the data records.

To determine how personal data should be protected using the GDPR, a clear definition of personal data according to the GDPR is needed. In article 4 of the GDPR [3], personal data is defined as:

"Personal data" means any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person.

Article 4 doesn't contain a specific list of attributes that need to be protected. The data officer has to decide which attributes have to be protected to comply with the article.

The GDPR distinguishes personal data and sensitive personal data. Sensitive data are defined in Article 9 as:

"Sensitive Personal Data" are personal data, revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership; data concerning health or sex life and sexual orientation; genetic data or biometric data. Data relating to criminal offences and convictions are addressed separately (as criminal law lies outside the EU's legislative competence).

A similar definition was present in the predecessor of the GDPR, the Data Protection Directive [9]. New in the definition of the GDPR are genetic data and biometric data [12]. New regulations allow sharing sensitive data for archiving, research or statistical purposes, but only when certain safeguards are in place. Data controllers should ensure data minimisation and use de-identification techniques on the data.

4.2 De-identification

Attributes in a data set that could lead to the name or address of an individual are called identifiers. De-identification techniques ensure that these identifiers are removed from the records. If the values are still needed for the research they are either replaced by values that no longer act as identifiers (anonymisation) or replaced by a pseudonym (pseudonymisation).

4.2.1 Anonymisation

Anonymisation is a de-identification technique where identifiers are either completely removed or replaced by more generalised or randomised values. The downside of anonymisation regarding research data is that the original values can't be restored. Therefore it is not useful when the data has to be reusable.

4.2.2 Pseudonymisation

Pseudonymisation replaces the identifiers by pseudonyms and stores the original values separately. With the combination of the data set and the separately stored values, the original link can be restored. Where data according to the regulations in the DPD are binary, they are either identifiable or anonymous, the GDPR covers pseudonymised data which are neither. The new regulations allow subsequent processing of pseudonymised data under certain circumstances (Article 6 [3]). Subsequent processing is the use of data for other purposes than it was originally intended for at the time of collection. The further research has to be "compatible" with the original research [7]. The GDPR is even more lenient towards research with scientific, historical and statistical purposes. In these cases, subsequent research can be conducted without special conditions if the protection of the privacy of data subjects is safeguarded.

This stimulates cooperation between research groups but has also led to criticism. This more lenient stance towards sharing research data implies that the writers of the GDPR have a lot of confidence in the protection pseudonymisation offers. Criticists question the reliability of pseudonymisation and argue that the risks of re-identification are too high [13].

4.2.3 Encryption

While a good implementation of the de-identification techniques is resilient against most attempts at re-identification of the original values, no current combination of protective measures is able to ensure total avoidance. Therefore the data should also be encrypted using state of the art encryption techniques. An overview of state of the art encryption techniques along with their performance is given in the thesis written by Victor Preda, resulting from the same Bachelor project.

4.3 EU Working Party

The Data Protection Working Party has provided guidelines for efficient anonymisation and pseudonymisation techniques [11].

4.3.1 Anonymisation risks

Merely removing the attributes that identify data subjects is not enough to ensure anonymisation. According to the Working Party, an effective anonymisation technique prevents the following three risks associated with anonymisation:

1. *Singling out*: a specific data subject is singled out from a data set by isolating some or all records that identify the subject.
2. *Linkability*: where records in the data set can be linked to the same data subject, even without retracing the identity of the subject.
3. *Inference*: the possibility to deduce an attribute by using the values of other attributes. Taking these risks into consideration, the working party lists two approaches to anonymisation: randomisation and generalisation.

4.3.2 Randomisation

Randomisation removes the strong link between the data subject and the data. To remove this link but preserve the distribution of the research data, attributes can be linked to different subjects or noise can be added to get rid of the original values. This is an effective method to protect against inference because the changed data will not lead to successful deductions. By adding noise the attribute can still be linked to the data subject but it will have an incorrect value. Assigning an attribute to a different data subject will get rid of the direct link, but will lead to a link to incorrect data. Linking subjects to incorrect values may be even more harmful than direct linkability given the circumstances.

4.3.3 Generalisation

Generalisation changes the scale of the data making the personal data less specific. An example is to change a city to a region or an age to a certain age range. This is effective against singling out because it puts the data subject in a group of people having identical values for certain attributes. It has to be carefully applied to protect against linkability and inference. Unlinkability is not guaranteed because it is still possible to link the records of groups of users from multiple data sets and find two corresponding records within the group. Inference is not prevented at all since you can easily obtain the value when you know which group a data subject belongs to.

4.3.4 Conclusion of the EU Working Party

A combination of randomisation and generalisation is the best way to anonymise data. But no combination of anonymisation techniques guarantees that data are protected against all three security risks. The data officer should therefore decide which risks can be taken considering a certain project. This is part of the 'data protection by design' principle which means that the assigned data officer should take privacy into consideration during each step of the research process and make tailor-made decisions about the collection and processing of personal data.

4.3.5 Pseudonymisation risks

Pseudonymisation can be seen as a variant of anonymisation, where specific attributes are replaced by pseudonyms. Because the data are transformed and not removed it is less secure than anonymisation. To show the difference in protection between anonymisation and pseudonymisation, the security risks that need to be prevented to ensure anonymisation are also evaluated for pseudonymised data.

It is still possible to single out a data subject since he/she is identifiable by a unique attribute, the pseudonymised attribute. The only way to guarantee unlinkability using pseudonymisation is by replacing all attributes that could identify a subject, removing all links to the original values and deleting the original data. Inference is still possible by analysing the attributes that correspond to a certain pseudonymised identifier whether in the same data set or across data sets.

4.3.6 Shortcomings of pseudonymisation

An example of the shortcomings of pseudonymisation is given by the Working Party. Consider a dataset where the data subject's Body Mass Index (BMI) is linked to the time they have received special assistance benefit payments (see Figure 4.1). A data subject's BMI is considered health data and is thus (as of Article 9 of the GDPR) sensitive data. The name, address and date of birth have been removed from the table and were replaced by an identifier generated by a hash function. As an extra protective measure, generalisation was used on the period of assistance benefit payments by counting only the years instead of the days.

Figure 4.1: An example of pseudonymisation using a hashing function that can be reversed

1. Name, address date of birth	2. Period of Special Assistance Benefit.	3. Body mass index	6. Research cohort reference no.
[redacted]	< 2 years	15	QA5FRD4
[redacted]	> 5 years	14	2B48HFG
[redacted]	< 2 years	16	RC3URPQ
[redacted]	> 5 years	18	SD289K9
[redacted]	< 2 years	20	5E1FL7Q

Aforementioned techniques have been successfully applied and therefore it is expected that this protects the privacy of the data subjects. However, if a third party knows the name, address, data of birth and the hashing function used the pseudonym can still be generated. Whether this is a risk that can be taken depends on the situation.

Another shortcoming of pseudonymisation lies in the fact that location data can be traced back to an individual. Researchers at the MIT [14] analysed a data set with the location data of 1,5 million data subjects in a 100 km radius and were able to single-out 95% of the subjects by using four location points. Using only two location points already 50 % of the data subjects could be singled out, probably by identifying their home and work locations.

4.4 Re-identification techniques and profiling

De-identification using pseudymisation or anonymisation does not guarantee the privacy of the data subjects under all circumstances. Anonymisation is a misleading term since data can never be fully anonymised (for more information, see Bijl's part of the project). The term suggests that every option to identify the data subject has been removed and all data can be safely published without the risk of identifying data subjects.

The same risks apply to pseudonymised data along with additional risks exclusive to pseudonymisation. These additional risks are related to the retrieval of the pseudonym by reversing the used algorithm or combining the pseudonym with the original data when they both have been compromised. These risks are described for each pseudonymisation technique in the results chapter. Medical data are the most sensitive data described in the DPD. The GDPR added more types of data to this highest risk category.

It is therefore interesting to see how medical facilities use pseudonymisation since they have a head start. Furthermore, a lot of medical data are very hard to anonymise because the real data is needed for analysis. The only option remaining for protecting the privacy of the patients or data subjects is pseudonymisation. These data are vulnerable to the attacks described in the following paragraphs.

4.4.1 Family attacks

An example of identifying a data subject using anonymised or pseudonymised data is shown by analysing research using genomic data. The data used in the deCode project [15] are collected by asking physicians to send data of individuals suffering from a particular disease to the Data Protection Commission (DPC) of Iceland. All identifying data except the social security number are removed by the DPC and the social security number is protected using a reversible encryption function. The DPC acts as a trusted third party between the physicians and the researchers and all data transfers are performed via the DPC.

The deCode project uses historical and genealogical repositories to find interesting patterns and useful patients. With these genealogies family-disease structures can be built, containing the gender and disease status (boolean: diagnosed with a particular disease or not). Comparing these structures with publicly available genealogical records can identify individuals.

4.4.2 Trail attacks

Trail attacks combine data from different hospitals to identify an individual. A patient's clinical and/or discharge data are often released as identified information or de-identified information that can easily be re-identified by linking it to public records. Since locations release this information independently, two tracks can be formed. The first track consists of the locations a DNA sample was left behind at and the second trail of the locations an individual has visited. These tracks can be combined to form the data trail of the genomic data (see Figure 4.2). Algorithms have been developed to combine these two tracks even when there is no perfect match between the two sets of data [15].

Figure 4.2: The data releases from locations I_1 to I_3 and the tracks created from these tables

I_1

τ^+	τ^-	
Name	*Pseudonym*	*DNA*
John	1G09JU3R	acag…t
Mary	F4P02SD4	accg…a

I_2

τ^+	τ^-	
Name	*Pseudonym*	*DNA*
John	4FG5097H	acag…t
Bob	U89KM32J	cttg…a

I_3

τ^+	τ^-	
Name	*Pseudonym*	*DNA*
Kate	AOEHA120	atcg…t
Bob	1X3C5VK4	cttg…a
Mary		

Identified Track

Name	I_1	I_2	I_3
John	**1**	**1**	**0**
Mary	**1**	**0**	**1**
Bob	**0**	**1**	**1**
Kate	**0**	**0**	**1**

DNA Track

DNA	Pseudonyms	I_1	I_2	I_3
acag…t	1G09JU3R 4FG5097H	**1**	**1**	**0**
accg…a	F4P02SD4	**1**	**0**	**0**
cttg…a	U89KM32J 1X3C5VK4	**0**	**1**	**1**
atcg…t	AOEHA120	**0**	**0**	**1**

4.4.3 High-level inference attacks

High-level inference attacks use domain knowledge to re-identify individuals. Generalisation is often used to protect fields that are not pseudonymised. For example, date of birth can be generalised to age. Attributes that are often generalised can be easily linked by their trivial connections. Consider two data sets: *Health*{name, address, birthdate, gender, zip code, hospital visit date, diagnosis, treatment} and *DNA*{age, gender, hospital visit date, DNA}. Combining these, the tuples {<birthdate, age>, <gender, gender>, <hospital visit date, hospital visit date>} can be made.

Using specific domain knowledge, even more links between data sets are possible. For example, records of medical prescriptions can reveal the diseases an individual is diagnosed with. Furthermore, a significant number of diseases can be linked to mutations in the genome. These diseases are listed in the International Classification of Disease Code - version 9 (ICD-9). These are just two examples of the sensitivity of genome data and the use of domain knowledge to identify individuals.

4.4.4 Low-level inference attacks

Low-level inference attacks are similar to high-level inference attacks, but the identification depends on multiple tuples instead of the single tuple in case of high-level inference attacks. Identification is done by drawing a conclusion from a combination of different tuples.

4.4.5 Lifelines

The danger of inference attacks shows why genetic data has been added to the "Sensitive data" category in the GDPR. Anonymisation doesn't help here because it would alter the genome data, which are needed for research. This is the case for a lot of medical data, making these vulnerable to above-mentioned techniques.

These vulnerabilities are also a concern in the "Lifelines" project conducted at the UMCG. Researchers collect data (urine-, blood- and hairsamples) from three generations of the same family. They are researching why some people fall ill and others don't. Oostergo told me that they use workspaces to access the data. The data are stored at one central location and researchers can only access it through a connection using Citrix Receiver. This circumvents data leaks that are caused by stolen USB sticks or laptops. The amount of researchers that can enter data into the system is very limited. According to Oostergo more researchers want to enter their own data since they are used to entering data themselves, but centralising data and using secure authentication ensures data protection through data minimisation.

4.5 Research focus

While the GDPR regulations cover all institutions storing personal data, this thesis specifically focuses on the protection of research data. Considering a data set collected during research, it evaluates the possibilities to protect the privacy of the data subjects according to the new regulations.

The main section of the thesis delves into the technical implementation of pseudonymisation. It compares the techniques that can be used to pseudonymise data. The list of these techniques is taken from the Article 29 [11], written by an EU Working Party providing advice on the technical implementation of the regulations stated in the GDPR. This section ends with the first research question of the thesis:

How does pseudonymisation help accomplish the privacy requirements as described in the GDPR and how can it be implemented to protect the privacy of data subjects in a set of research data?

In addition to the technical implementations of pseudonymisation, this thesis focuses on the risks that remain when data has been pseudonymised. The various de-identification techniques have shown that data can still be traced back to an individual data subject when it has been pseudonymised. Considering these security risks, pseudonymisation might not be as secure as the recommendations in the GDPR suggest. Comparing all benefits and risks will lead to the main research question:

Is pseudonymisation the silver bullet for compliance with the GDPR?

Chapter 5

Methods

The attributes that act as identifiers as mentioned in the GDPR are analysed to see in which way these attributes can compromise the privacy of the data subjects and whether all compromising data attributes are present on the list. This is done by comparing the list to the one present in the Health Insurance Portability and Accountability Act (HIPAA), data protection regulations for health data in the United States.

A comparative analysis of pseudonymisation techniques shows the strengths and weaknesses of the various techniques. An exciting new technique gives a glimpse into the future of pseudonymisation.

A conversation with René Oostergo, Lead Engineer at the Trial Coordination Center of the University Medical Center Groningen (UMCG) gives more insight into the problems faced while protecting medical data and how to address these using anonymisation and pseudonymisation techniques. It is interesting to consult people working at medical facilities since medical data have been protected with anonymisation and pseudonymisation techniques to comply with the predecessor to the GDPR, the DPD [9]. The GDPR now requires other personal data to be protected using these techniques and it is useful to learn from people who already have experience using these techniques.

After analysing how pseudonymisation can be used to protect personal data according to the GDPR, the dangers that accompany the use of pseudonymisation are discussed. The presence of efficient re-identification techniques might make using pseudonymisation less secure than the GDPR claims. An analysis of articles in the GDPR combined with Oostergo's domain knowledge illustrate how security risks caused by human factors can be limited.

Chapter 6

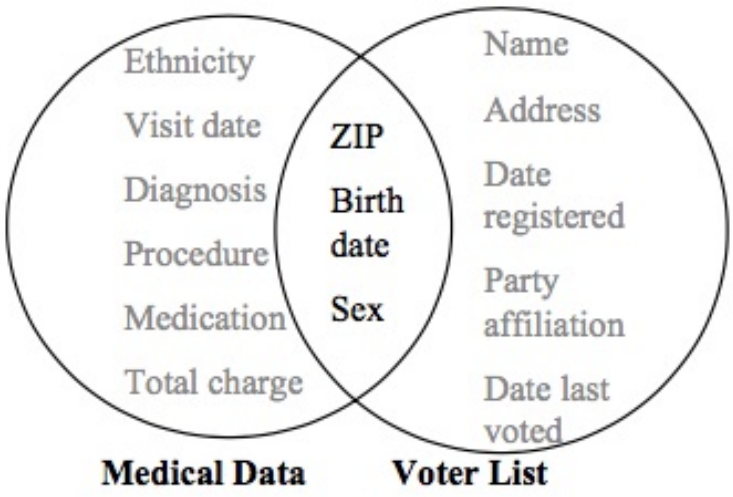
Results

6.1 Comparison of GDPR and HIPAA

Since the United States has recently introduced strict regulations for personal data, it is interesting to compare the GDPR to the Health Insurance Portability and Accountability Act (HIPAA) introduced in 2012 [16]. The Safe Harbor method in the HIPAA mentions all fields that reference the data subject or relatives, employers or household members of the data subject that therefore need to be de-identificated.

The GDPR and the HIPAA both mention names and location data as identifiers. The HIPAA defines location data more clearly as every geographical subdivision smaller than a state. The term location data implies that GPS data is also included in this category, impacting a lot of research projects using location indicators. Dates related to the data subject are outlined in the HIPAA but are missing in the GDPR article apart from the ambiguous 'social identity' category. Using de-identification on dates of birth and ZIP codes is very important since Sweeney (1990) has shown that 87% of the population of the United States could be identified using only names, ZIP codes and gender [17]. When the experiment was conducted in 1990, medical records contained all personal details except names and addresses. Cross-referencing these medical records against openly available voter registrations using the mutual ZIP codes, birth dates and gender revealed names and addresses of patients.

Figure 6.1: The overlap used by Sweeney in her experiment



6.2 Pseudonymisation techniques

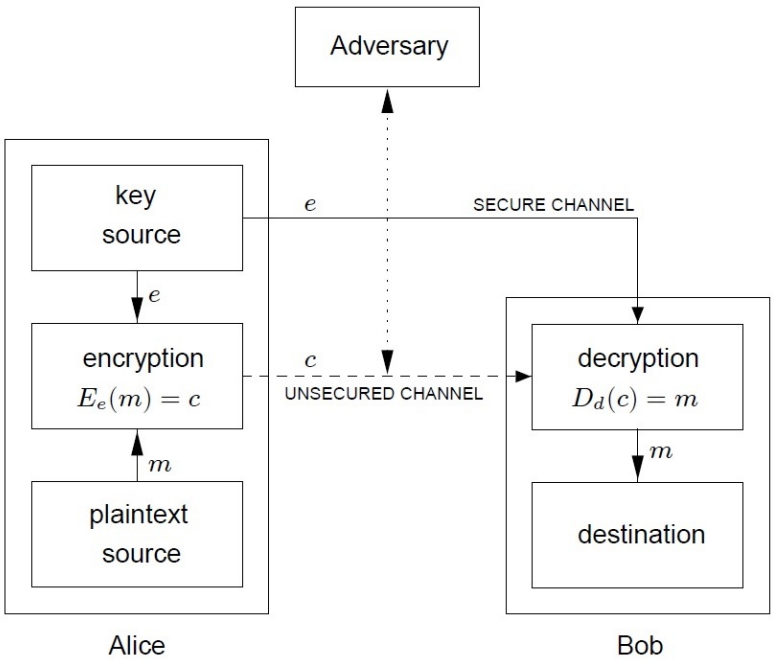
The EU Working Party has listed some pseudonymisation techniques that can be used to comply with the GDPR [11]. They can be divided into three categories: encryption, hashing and tokenisation.

6.2.1 Encryption with a shared key

Same key encryption is an encryption scheme using the following transitions: Encryption using $\{E_e : e \in K\}$ and decryption using $\{D_d : d \in K\}$ where K is the key space [18]. Since the decryption scheme is known when the encryption scheme is available, this is called encryption using a symmetric key.

The encryption key e is a permutation on alphabet A. To encrypt a message this permutation is applied on blocks of a fixed number of characters. To decrypt the message the inverse $d = e^{-1}$ is used to retrieve the original message. The communication between two people using the encrypted message and keys can be seen in Figure 6.2. It is assumed that both people know the encryption transformations. The encryption key e is shared between sender and receiver. The reverse relation between d and e means that d has to be kept secret as well, since d can be obtained from e. The receiving party constructs the decryption key d from e and uses it alongside the publicly known encryption scheme to decrypt c and retrieve the original message.

Figure 6.2: Encryption with the use of the same key



This encryption technique can be used for pseudonymisation by replacing the identifiers in the original data set by pseudonyms generated using an encryption scheme and key e. This encryption key e has to be stored at a separate secure location. If the key is not kept separately, a security breach will yield both the data and the key rendering the whole pseudonymisation process useless.

6.2.2 Unkeyed Hash function

A hash function uses an efficient algorithm to map binary strings with an arbitrary length to binary values of fixed length, the hash value. The probability that an arbitrary string gets mapped to a n-bit hash value is 2^{-n}. A hash function is chosen such that it is computationally infeasible to find two inputs that are hashed to the same hash value. Another requirement of the given hash function is that it is implemented such that given a hash value y it is computationally infeasible

to compute the corresponding input x. The MD5 hashing technique has been outdated for years and does not provide enough protection. The first security breaches on the SHA-1 technique have surfaced [19]. A hash function from the SHA-2 and SHA-3 families can be used as a secure hashing technique.

Hashing functions can efficiently be computed and therefore are an accessible technique to use for pseudonymisation even for smaller companies [11]. The downside is that these functions are prone to brute-force attacks. Salted hashing can be used as an extra protective measure. This adds a random value to the attribute that is hashed. This provides protection against attacks where a hash table with commonly used attributes is used to retrieve the original data. These Pre-computed lists of hashes can't be used because the added salt is a random value. The original value can however still be derived using a dictionary attack or brute-force attack, especially if the data type of the attribute is known [20]. For example, a social security number has a certain length which limits the options for the value of the original data. If this length is known and the salt is known, a brute-force attack can still retrieve the original value.

6.2.3 Keyed hash function with stored key

This is a special case of a hashing function where the hash is protected by a secret key as input. This secret key has to be stored separately to ensure that attackers do not have access to it when the hash is compromised. The secret key is shared with the receiving party over a secure channel. When the receiving party wants to retrieve the original data, the hash function is replayed with the secret key as input. Without knowing the key, the number of possibilities to be tested increases substantially. The difference between this secret key and a salt is that the salted part is not secret.

6.2.4 Deterministic encryption

Deterministic encryption is an encryption method that returns the same cyphertext with each iteration of the algorithm while using the same plaintext and key. An example of deterministic encryption is RSA, a public key encryption technique [21].

The description of the algorithm in the article of the Working Party suggests that the encryption technique contains a probabilistic element. Where deterministic encryption algorithms produce the same cyphertext each time the same plain text is encrypted, probabilistic encryption algorithms produce a different result with each execution. This protects the data against brute-force attacks using the public encryption key. This can be achieved by applying the Optimal Asymmetric Encryption Padding (OAEP) scheme on the plaintext before encryption.

6.2.5 Tokenisation

Tokenisation is a new technique, often used for protecting bank transactions [22]. When certain attributes are tokenised, three different steps are performed:

1. A token is generated to replace the original data. A token is a random string with no mathematical relationship with the original data and therefore can't be reversed.
2. The token server encrypts the original data using a key that is only known within the server.
3. The original data and the generated token are stored together in the token database.
4. The token is returned to the calling application. The calling application removes the identifiable data and replaces it with the token.

It has been adopted by credit card companies because it has two major benefits over encryption and hashing:

1. A token created by tokenisation can't be restored to its original value. While encryption techniques and hash functions obfuscate the data, tokenisation removes the original data and replaces it by a token. The pseudonyms that replace the original values in the case of encryption or hashing are gained by using a certain mathematical formula. If a third party is able to recreate this mathematical formula, the original values can be retrieved. Because tokenisation does not use any mathematical basis for the replacement this is not an option.

2. A token maintains the structure and data type of the original value. It makes it less prone to attacks, because attackers will think that these are already the original values.

Since tokenisation is already in use to protect bank transactions, VISA has provided guidelines for secure tokenisation. While tokens can be generated in various ways, it is advised to use random number generation to generate the tokens that replace the original data. Random numbers are simple to generate and there is no mathematical function used that can be reversed to retrieve the original values. The original values are only retrieved by looking them up in the token server database.

A token server database is a database that holds the links between tokens and their original values. A company can decide to host their own token server database or use the database of a Trusted Third Party (TTP). Oosterloo told me that the UMCG relies on a TTP for the storage of the pseudonyms used in a project called Parelsnoer in which medical research data is shared between eight University Medical Centres. This caused problems when the DigiNotar hack [23] compromised the TTP. It took three years to rebuild the token storage at another company, since there were no standards for storing tokens. During this period no information was shared between the participating University Medical Centres.

6.2.6 Comparison of techniques

For a comparison of encryption and tokenisation see Figure 6.3 [24]. Tokenisation is better for structured data within databases, since there is no mathematical connection between the pseudonym and the original data. It is stored at a single location, making it easier to implement data minimisation. However, it can't be used for files and unstructured data.

Encryption is a convenient technique for data transfers and even when hashing or tokenisation are used as pseudonymisation techniques, encryption is still used to transfer the files. The receiver can retrieve the original data using the pseudonymised data and the encryption key. Tokens can only be re-identified by looking them up in the token database and therefore have to be encrypted at the token server before transferring it. Since hashing is a one-way function, hashed pseudonyms also have to be encrypted at the hash storage before sending them.

Figure 6.3: A comparison of encryption and tokenisation

Encryption	Tokenization
Mathematically transforms plain text into cipher text using an encryption algorithm and key	Randomly generates a token value for plain text and stores the mapping in a database
Scales to large data volumes with just the use of a small encryption key to decrypt data	Difficult to scale securely and maintain performance as database increases in size
Used for structured fields, as well as unstructured data such as entire files	Used for structured data fields such as payment card or Social Security numbers
Ideal for exchanging sensitive data with third parties who have the encryption key	Difficult to exchange data since it requires direct access to a token vault mapping token values
Format-preserving encryption schemes come with a tradeoff of lower strength	Format can be maintained without any diminished strength of the security
Original data leaves the organization, but in encrypted form	Original data never leaves the organization, satisfying certain compliance requirements

Hashing is not included in the comparison but can be used as a less expensive method to secure data. While tokenisation is more secure because it doesn't depend on mathematical principles and encryption is used for sharing data, hashing can be a useful technique for smaller projects where more risks can be taken. Computing hashes is cheap and it is safe when a modern hashing technique is used.

While research projects vary in size and budget, this comparison shows that there are pseudonymisation techniques for various budgets and purposes. When implemented correctly using state of the art algorithms, they secure personal data within an acceptable margin of risk that is inevitable. The EU Working Group encourages experimentation with new techniques as long as certain safeguards are in place. [11].

6.2.7 In development: polymorphic encryption and pseudonymisation (PEP)

A new technique is in development at the Radboud University: polymorphic encryption and pseudonymisation (PEP) [25][26]. Bart Jacobs, Professor of Software Security and Correctness at the Radboud University, is leading the team that built this new solution. It is currently in pilot phase in which RadboudUMC, the university hospital, and Verily are using it for joint Parkinson research. Verily is a life sciences company under the flag of Alphabet and besides Google. The data of 650 Parkinson patients is analysed to be able to provide personalised treatments based on statistical outcomes of large scale analysis of patient data.

Due to the large amount of personal data that is needed to conduct such research, this is an excellent test case for the new PEP technique. The Digital Security group at the Radboud University is providing the PEP framework. It will be published as an open-source project after the pilot phase and will be deployed at external partners. According to Oostergo, the UMCG has shown interest in using this new technique when it is available, . The pilot began in May 2017 and is expected to run until October 2021. In June 2019, the database has to be fully operational with data from all 650 patients and possibly open for collaborations with other research groups.

6.2.8 Polymorphic encryption

In traditional public key encryption, only the holder of the private key can decrypt the data. For collaborative projects like the Parkison research, many people would like to have access to the data. Sharing the key with all these people undermines the security of the data. PEP uses malleable locks (see Figure 6.4) which can be opened by several keys. An third party called the TransCryptor can change the lock in such a way that the right person can access it with his/her key by a process called re-keying. This re-keying is done by the TransCryptor without it having access to the data in the process.

Figure 6.4: An illustration of malleable locks

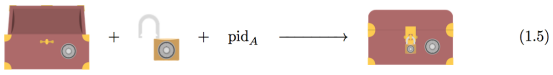
6.2.9 Polymorphic pseudonymisation

Each patient gets a unique identifier pid_a. This identifier is morphed into pseudonyms by the central TransCryptor that differ per data handler. The pseudonym for data handler X, generated from pid_a is called the local pseudonym of pid_a at X. These local pseudonyms are created by the TransCryptor in a blind matter, without learning anything about the data. Doctors store the link between the identity of a patient and the local pseudonym. Researchers use only the local pseudonyms to organise the data. Pseudonyms have two wheels as described in the previous paragraph. One lock to make the pseudonym polymorphic and one lock to change the data without accessing it. A polymorphic pseudonym is formed by storing pid_a and a polymorphic lock together (see Figure 6.5). This serves two purposes:

1. A local pseudonym $pid_a@B$ (the local pseudonym used by data handler B) can now be formed by turning the wheel of the polymorphic pseudonym to position B using a process called re-shuffling.

2. If the wheel on the pseudomorphic lock inside the pseudonym is turned to $pid_a@B$, B can open the chest and retrieve the local pseudonym $pid_a@B$.

Figure 6.5: Creating a polymorphic pseudonym



When data have to be sent to the central server, the data and the polymorphic pseudonym are sent to the Transcryptor. The Transcryptor does not touch the data but turns the locks on the pseudonym to the position S related to the central server. The central server opens the locks and retrieves the local pseudonym $pid_a@S$ and uses this as database key to store the data.

Retrieving data from the central server works in a similar way. If doctor B needs data on patient A, the identifier pid_a is sent to the Transcryptor. The Transcryptor turns the locks on the pseudonym to the position of the central server. The central server can now open the lock and retrieve the local pseudonym $pid_a@S$. The server gets the data from the database and returns the locked chest. The Transcryptor turns the locks to position B and doctor B can access the files.

6.3 Human influence

The techniques described so far only protect the privacy of data subjects when they are put into practice by the people handling the data. To do so the consent of data subjects is needed as well as the cooperation of the personnel.

6.3.1 Consent

An important factor in the GDPR is consent. Before data from a certain data subject can be processed, the subject has to give his/her consent. For personal data, subjects have to give "unambiguous consent". Lawyers will have to find out what "unambiguous consent" exactly entails when the regulations are enforced. For sensitive personal data a subject has to give explicit consent. This is defined in Article 29 of the Working Party [11] as:

"All situations where individuals are presented with a proposal to agree or disagree to a particular use or disclosure of their personal information and they respond actively to the question, orally or in writing."

Furthermore, Article 7 of the GDPR states that a subject should have the possibility to withdraw consent and withdrawing consent should be as easy as giving it. This should be part of the 'data protection by design philosophy', stating that every system that holds personal data is built with privacy in mind from the start of the project. Computer scientists should build their systems in such a way that data can be easily removed when data subjects ask them to.

6.3.2 Compliance

Oostergo stressed that the compliance of personnel is very important in introducing new regulations in a corporation. This can be done by providing training sessions. Most companies organise these training sessions, but only involve new personnel. While these new employees oftentimes will use the new techniques taught, the long-time employees stick to the techniques they are used to and require a lot of convincing to change their behaviour. Oostergo mentioned that the UMCG needs a lot of training to be ready for the GDPR regulations. Many long-time employees only protect their data subjects by encoding them without an encryption method, merely replacing

the names and addresses by patient codes. Even more exemplary, the tables with the encodings of these data subjects are often stored on the same data drive as the research data in an Excel file. They should be stored on a separate drives but long-time employees don't see the risk involved.

While it is important to build a safe environment to store sensitive data, it is even more important to incentivise the personnel to use these new techniques. This should not be limited to new personnel; long-time employees should also be informed on the new regulations and the consequences for their tasks.

Chapter 7

Discussion

After evaluating anonymisation and pseudonymisation techniques and their weaknesses, it is apparent that full protection against re-identification attempts is not possible. The data officer has to evaluate which risks can be taken in the project he/she is working on. The overarching protection goal of data minimisation has to be taken into account. The data officer should always limit the number of data that are collected and the amount of people that have access to the data.

For sensitive data this can be best achieved by using a central location to store the data. Organisations can build a server themselves or rely on a TTP. The researchers can access this using remote workspace environments. The data officer should use state of the art authentication techniques to make these connections secure and make sure that only authorised personnel has access to the data. Even when data is limited to these central servers, researchers should be aware that they can't publish all data they consider to be anonymous. For example, genomic data can be traced back to the individual who provided it.

There will always be risks involved in dealing with personal data. The fact that human processors have to access the data already exposes the data minimally by displaying them on a screen. Co-workers can see the data when their colleague is examining them. The data are even more exposed when the they are shared among a lot of researchers. This is why the data have to be limited to the processors that really need to have access to the data to fulfil the requirements of the research.

The GDPR is written to be future proof and therefore the technical terminology is left intentionally vague. If the writers would have stated specific techniques to use for data protection, the document would be outdated very quickly. Instead they require data officers to use state of the art techniques. While the EU Working Party has given advice on which techniques to use at the moment, these techniques may be considered insecure in the future. New pseudonymisation techniques have to be analysed to test their safety and older techniques have to be constantly re-evaluated. The PEP technique looks promising and has to be revisited when the researchers at the Radboud University have made progress in the pilot phase.

7.1 Is pseudonymisation the silver bullet?

Pseudonymisation has been brought forward as the most important new technique in the GDPR. The GDPR even allows sharing of data when they have been pseudonymised. Section 1.4 has shown that pseudonymised data can be re-identificated and can't be considered as anonymous data. Pseudonymised data are still personal data and therefore the data controller still needs the consent of the data subjects [13].

However, pseudonymisation greatly improves the security and it should still be used if anonymisation is not an option [27]. If the importance of pseudonymisaton is downplayed and anonymisation is not an option because the identity of the data subjects has to be reversible, data controllers may be incentivised to use no data protection apart from encryption. If anonymisation is not an option, pseudonymisation is a good alternative because it removes the direct link between the

data subject and the data. If the table with the connection between the data subjects and the pseudonyms is stored at a different location, two locations have to be compromised to be able to directly identify the data.

7.2 Future work

It would be interesting to compare this analysis to upcoming documents which provide advice on dealing with the GDPR from a technical perspective, such as the missing chapters of the SDM. It would also be interesting to see whether this analysis is useful to Esther Hoorn and people from various faculties of the RUG who are analysing the GDPR from their perspective. Tools are being built for researchers within the RUG to comply with the GDPR and this analysis might be helpful.

Chapter 8

Project Discussion

The General Data Protection Regulation (GDPR) protects the privacy of European citizens by limiting the storage and movement of their personal data. The Standard Data Protection Model (SDM) provides measures to ensure protection goals of privacy. In this project three technical measures or implementations were extracted from the SDM to evaluate how these can be used in the scope of the GDPR. The technical implementations are encryption, anonymisation and pseudonymisation and were evaluated by each individual author of this document. Their findings are summarised below, for a more elaborate discussion of those measures we refer to the individual theses conducted by each author.

8.1 Encryption

In order to achieve data security, encryption is one of the proposed measures in the GDPR. In Preda (2017) cryptographic techniques were presented and analyzed in order to provide a more technical overview of this proposition. From a legal perspective, all the presented methods qualify as solutions for securing data. From a technical perspective, however, more details are necessary. Thus, two categories of cryptographic methods were described in this paper: classic encryption techniques and multimedia encryption techniques.

Following the theoretical analysis, it was concluded that the current standard for encryption (Advanced Encryption Standard) is the most balanced solution. Theoretically, all data connected to a research project, which deals with personal details of human subjects, has to be secured using this algorithm. The full implementation of the algorithm is considered secure and efficient. Practically, in order to further strengthen the security of the data and at the same time offer a high level of speed and efficiency, alternative methods should be considered.

Implementations of classical methods on specialized hardware, such as Field Programmable Gate Arrays, should be considered. This provides a considerable increase in performance when manipulating data. Furthermore, specialized hardware such as the one previously mentioned, could be integrated in order to provide external storage. Further research is needed, however, in order to transform this idea into a practical one.

Given the diversity of research projects, an analysis of each individual project should be conducted in order to determine the most efficient methods to be used for aforementioned project. If the project is composed of other types of files than text files, specialized methods should be employed. The presented methods are, as the classical methods, considered to be secure. Furthermore, by taking advantage of the formats of the files they provided a higher efficiency for multimedia files than classical methods can be achieved. Moreover, by using different methods for smaller parts of a project the overall security of the project is increased in case one method is compromised.

Encryption combined with the following two techniques offers a basis for security systems which are in accordance with the GDPR requirements. Furthermore, if detailed analysis is performed

for individual projects and these methods are adapted, the requirements of the GDPR can be surpassed and higher security can be obtained.

8.2 Anonymisation

Anonymisation techniques are considered very promising for publishment of data. How anonymisation relates to the GDPR was evaluated by the thesis about data anonymisation. To summarise the findings of Bijl (2017) the following four parts are discussed. First the GDPR and its effects for anonymisation were examined. Second former studies on anonymisation that result in different opinions on anonymisation are shown. Third real state of the art techniques to anonymise data are evaluated. Last the conclusions and findings of this thesis are provided.

8.2.1 General Data Protection Regulation and anonymisation

The GDPR focuses on the protection of personal data of European citizens. As stated in the GDPR, personal data is any information where individuals (data subjects) can be identified from. Information from which no data subject can be identified is regarded as anonymous. As a result, the GDPR plays no role while dealing with anonymous information. This fact enables data controllers to publish data from which no individual can be identified.

8.2.2 Former studies

In this thesis different studies on anonymisation are shown and evaluated. Sweeney (2002) [17] has shown that only a combination of gender, date of birth and ZIP code can identify 87% of citizens in the United Stated. Narayanan and Shmatikov [28] applied a new de-identification technique on an 'anonymous' data set published by Netflix. They were able to identify an enormous number of users by combining these Netflix users with users of the Internet Movie Database (IMDB). These examples showed that removing direct identifiers such as name, username, etc. is insufficient to anonymise data. Ohm [29] stated that true or proper anonymisation can not be reached when publishing data with utility. Ohm provided a trade-off between data utility and privacy where both dimensions cannot be fully reached. Therefore data with utility cannot have zero risk of reidentification. The question that arises is what privacy guarantees of data are enough?

8.2.3 Anonymisation techniques

Four privacy models to anonymise data were evaluated in this thesis. The models are: k-anonymity, l-diversity, t-closeness and differential privacy. These were evaluated to examine which privacy guarantees each model provides.

k-Anonymity is a model that changes the quasi-identifiers in such a manner that each equivalence class has at least k rows. When k-anonymity is satisfied, within each equivalence class an individual cannot be distinguished from at least k-1 other individuals. By knowing the quasi-identifiers an adversary is able to know the equivalence class of an individual but still has a probability of 1/k to learn the exact row in that equivalence class. k-Anonymity is still vulnerable to the homogeneity and the background knowledge attack.

l-Diversity is a model that must satisfy variety of sensitive values within every equivalence class. Every equivalence class must have at least l different sensitive values for each class. The drawbacks of l-diversity are the skewness and similarity attacks.

t-Closeness is a property that ensures that every equivalence class and the overall distribution have no significant difference. The difference between each class and the distribution as a whole cannot be higher than a threshold t. t-Closeness is vulnerable to attribute linkage and similarity attacks.

Differential privacy is a different privacy model that satisfies a difference of exactly one record in every published table. As a result the overall conclusions will roughly stay equivalent, but

randomisation provides uncertainty about the link between the data subject and the data itself. Problems of differential privacy can be restrictions on queries and the value of ϵ.

8.2.4 Conclusions

In this thesis it was shown by literature that proper anonymisation is practically impossible to achieve. Therefore choices must be made regarding data privacy and utility. In the evaluation it is shown that each privacy model provides privacy guarantees but are still vulnerable. The GDPR does not specify strict properties or values for these privacy models. From here it can be concluded that data controllers must show effort in the direction of anonymisation.

8.3 Pseudonymisation

The GDPR defines pseudonymization as "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information" [3]. When the link between research data and the data subjects involved has to be maintained, pseudonymisation can be used as a protective measure. All direct and indirect identifiers are replaced by a pseudonym using one of the pseudonymisation techniques. Direct identifiers are attributes that can identify an individual without the use of additional data (name, address, social security number etc.). Indirect identifiers can identify an individual by combining them with other attributes from the data set. Obtaining the identity of the data subject from pseudonymised data is done by using the de-identification technique corresponding with the pseudonymisation technique used.

Pseudonymisation is important in the light of the GDPR because it provides some benefits while processing non-anonymous data. Where the regulations in the DPD are binary, data are either identifiable or anonymous, the GDPR covers pseudonymised data which are neither. The new regulations allow subsequent processing of pseudonymised data under certain circumstances (Article 6 [3]). Subsequent processing is the use of data for other purposes than it was originally intended for at the time of collection. The further research has to be "compatible" with the original research [7]. The GDPR is even more lenient towards research with scientific, historical and statistical purposes. In these cases, subsequent research can be conducted without special conditions if the protection of the privacy of data subjects is safeguarded.

The EU Working Party has suggested some adequate pseudonymisation techniques. These can be divided into three categories: encryption, hashing and tokenisation.

8.3.1 Encryption

When using encryption as pseudonymisation technique, the identifiers are encrypted and the resulting ciphertext acts as the pseudonym. The key that is generated by using the encryption scheme has to be kept at a separate location. When a receiving party wants to identify the data, the key has to be sent over a secure channel. The key is used, combined with the publicly known encryption scheme, to decipher the pseudonym. Because the data can be easily re-identified using the key, encryption is a good technique for sharing research data.

8.3.2 Hashing

Hashing functions are one-way functions that turn the data into a binary string of fixed length. Given the same input and hashing function they provide the same output. This makes it vulnerable to dictionary and rainbow attacks. Attackers generate the hashes of commonly used phrases and compare the hashes to the hashes they want to identify. Adding a random value before hashing (salting) or a secret key protects against some of these attacks.

Because hashing functions are one-way functions the identity can only be recovered by retrieving the combination of the original data and the hash. These combinations have to be stored separately. Hashing can therefore not be used to transfer data. Tokenisation is a better technique for storing

data and encryption is better for sharing, but hash functions are easy to compute and can therefore be a good solution for smaller projects where budget is limited and more risks can be taken.

8.3.3 Tokenisation

Tokenisation replaces the identifiers by random values and stores the combination of the original data and the tokens in a token database. Where encryption and hash functions are based on mathematical principles, tokens are random making them less prone to brute-force attacks. It also preserves the data structure, this can be useful when a subset of the data is needed. For example, the digits of a ZIP code can be preserved while the letters are pseudonymised.

8.3.4 Trusted Third Party

The combinations of the pseudonyms and the data have to be stored separately. Research groups can decide to build their own system to store these combinations or rely on a Trusted Third Party (TTP). Relying on a TTP leads to a paradox, according to Jeanne Mifsud Bonnici, Professor in European Technology Law and Human Rights at the Law department at the University of Groningen:

The paradox is that the fulfilment of the legal obligation (data protection by design and by default) by the person responsible (data controller/customer of security products/systems) is heavily dependent on persons (manufacturers of security products/systems who design the products/systems) not subject to the data protection legislation [8].

From a law perspective it might seem strange to rely on a TTP for providing an important part of your system. However, it is very common to do this in computing science. It is very expensive to build a proprietary system and test it extensively before using it. We argue that relying on a TTP that can provide a system that is appropriate for the research conducted and has been tested thoroughly is more secure than using a new system.

Therefore, the accountability for these legal obligations should be divided. The TTP is accountable for providing secure storage while the data controller is accountable for overseeing the research. Even though the data are stored at a separate location, the data officer is still in charge of the legal obligations of the GDPR such as minimising the data collected, the processors involved etc.

8.4 Case study

To illustrate the findings in this project the authors perform a case study on an existing research. The research is called 'Engaging research participants to inform the ethical conduct of mobile imaging, pervasive sensing, and location tracking research' performed by Nebeker et al. [30].

8.4.1 Research description

In this study a number of willing participants were monitored for 7 days. This monitoring was performed using diverse devices. Each subject wore a digital camera on his/her chest, a GPS device, an accelerometer around the waist, and 2 more accelerometers on each wrist. The researchers of the iWatch study had as primary goal to monitor how the ethical framework is applied to this type of study. This ethical framework consisted of 4 main parts:

1. Informed written consent of participant.
2. Privacy and confidentiality
3. Non-maleficence
4. Autonomy of third parties

Out of these 4 parts only one is partially directed towards a technical aspect. In privacy and confidentiality it is mentioned by the research that:

Devices should be configured so that data can only be retrieved by the research team. It should be impossible for participants or third parties who find devices to access images.

This is dealt with by employing encryption. It is not specified however how this is done and what cryptographic method is employed. Anonymisation is partially implemented through the specifications of the Health Insurance Portability and Accountability Act (HIPAA). This however is not sufficient in order to provide proper security measures at the level required by the GDPR. The third measure in the scope of this paper, pseudonymisation, is non-existent in this framework and therefore also non-existent in the security measures adopted for the iWatch study.

It should be noted that project descriptions are more oriented towards an abstract description of the security measures employed in order to assure security. From a technical point of view details regarding the way in which data is processed, stored and manipulated are of high importance. The hardware which is used in these processes is also important. This is due to the fact that these play a key role in the technical analysis and decision making of which security methods should be employed and how they should be implemented.

8.4.2 Personal data

We have seen in the iWatch study, data is collected by multiple monitors to indicate the behavior of the participants. The monitors that were used are: a SenseCam imaging device, a Global Positioning System (GPS) tracking device and three activity monitors called accelerometers. We have seen in the General Data Protection Regulation (GDPR) personal data is defined as any information where people can be identified from. Moreover the GDPR specified in definition 4 of article 4 certain identifiers:

... in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Because of the above-mentioned description the data collected in the iWatch study falls under the protection of the GDPR. An example can be illustrated through location tracking technologies, these are used in the study to collect data. It is stated in the GDPR that data subjects can be identified by investigating their recorded location. Moreover, in order to monitor to a high degree of precision subjects' daily behavior, the SenseCam takes an average of 3000 images a day. This large amount of information could contribute to compromising the identify of the data subject.

The authors state that the data may be considered sensitive, and therefore adhered to the (American) HIPAA regulations. In the GDPR location data is not considered sensitive personal data, because it falls under personal data as can be seen in the aforementioned definition. However, demographic data is collected which fall under the sensitive personal data category because it contains data concerning 'racial or ethnic origin' [31]. To use sensitive personal data, researchers should have 'explicit consent'. Because the researchers explained to the data subjects where the data will be used for and collected formal consent, they have acted in line with the GDPR.

8.4.3 Discussion

Encryption

According to the authors of the study, the method of encryption is employed in order to secure images taken by the camera. This is done in order to prevent the subject and any other third party from interacting with the taken images. Despite this measure, a number of issues arise. One of them is the security of the other 4 devices worn by the subject. No security measure is mentioned about them. While in a person's mind, subjectively, images seem to be more of a security breach than the rest of the data, security measures should still be employed for the former. Valuable information, possibly even more exact information, can be extracted from GPS coordinates.

If, as a conceptual case, a malicious third party would be interested in determining different positions of a subject, GPS data would be far more valuable than images. This is because if one

would have access to precise GPS coordinates, then images would become irrelevant for this case. Because of the situation in which measuring devices are on the subject and not in a controlled environment, strong security measures should be implemented for each component. Depending on the capabilities of each piece of equipment, vulnerabilities could be exploited in the places that the subject is present. This could happen with or without the subject's knowledge.

Another issue of concern would be the fact that even though it is specified that encryption is used in order to secure images, it is not specified what type of encryption is used or how it is implemented in the device. These are technical details which are irrelevant for the purpose of the study. Yet they are highly important for a proper security analysis in order to ensure the safety of the sensitive data collected about the subject. These details should be provided in a special section of the research paper by the researcher or the data officer.

Furthermore, after retrieving the data from the devices, it is also not specified if any security measures are employed for the storage and processing of the data by the researchers. This is also one important aspect. For instance, when subjects are able to review the images, there could be an opportunity for exploitation. If no security measures are implemented, the fact that the images were encrypted on the device is irrelevant if the subject then has free access to them at the researcher's office. Depending on the methods in which the subject can interact with the images, multiple possibilities exist. These can range from copying the images at that moment to possibly inserting malicious code which could possibly run when it has access to all recorded images. This code could copy all of these and send them to a remote location to which a third party would have access. This would not only achieve the purpose of accessing one subject's data but, also accessing all subjects' data.

Moreover, depending on how these images are stored, the same method could also have access to the rest of the data (i.e. GPS coordinates and the other information regarding the data subject). In order to prevent this, close inspection of the way in which the subject is allowed to interact with the recorded data must be performed. Additionally, the methods in which the data are stored and further processed should also be analyzed. These possible vulnerabilities can be solved using encryption in relation with the following two techniques. The main suggestion for this project and others similar to it, is that in order to properly provide security, in accordance with the GDPR, multiple technical details must be inspected, recorded and made available.

Anonymisation

The iWatch study raised a couple of issues from an anonymisation perspective. Standards to improve privacy were applied from the Health Insurance Portability and Accountability Act (HIPAA). The authors' motivation to satisfy those requirements, was that the data may be considered sensitive, and as a result they opted the HIPAA standards. In the HIPAA, a so called safe harbor model that modifies or masks direct identifiers is suggested. The problem here is, as one can see in the individual thesis of anonymisation, that the HIPAA standards are considered insufficient to anonymise data. Earlier and well-known studies have illustrated the fact that individuals can be identified from a combination of quasi-identifiers which are preserved by the safe harbor mo del. Therefore applying the HIPAA standards is a way to improve anonymisation, although it is considered a poor method towards anonymisation.

Another issue of the iWatch study is shown by table 2 in the paper. The table shows several demographics of participants in the iWatch study such as the frequency of certain statistics in the data. One particular frequency is important from an anonymisation point of view. The frequency of the group American Indian/Native American has frequency 1.2% which equals one single person in the data. This is a problem because individuals can be identified from this dataset. If an adversary has background knowledge, such as knowing a American Indian/Native American that participated in this study, he is able to discover sensitive information of this person. This situation is highly undesirable and is caused by the fact that the table has the property of 1-anonymity. Therefore, individuals can be easily identified from this data set. To improve this, k-anonymity techniques can be applied to this data set.

Pseudonymisation

While pseudonymisation was not used by the researchers, it could improve the privacy of the data subjects in this context. All direct identifiers were removed in an attempt to make the data anonymous. Last paragraph has shown that the data are not anonymous and can even identify individual subjects. Because the group with subjects from American Indian or Native American origin contains only one subject, it would be desirable for the privacy of the data subjects to remove this attribute. It is not a direct identifier, since it can only directly identify an individual if you know a person from American Indian or Native American origin who participated in this research.

Removing the attribute from the data set is an option, but the researchers must have a purpose for these data in their research, otherwise they wouldn't have collected them. A better solution would be to pseudonymise the demographic data, replacing the demographic data by identifiers and storing the combination of pseudonym and demographic at a separate location, for example on a server of a TTP. After pseudonymising the data, the pseudonymised data can be used for the research and if researchers need to link their results to demographics they can request these identifiers from the TTP.

If pseudonymisation is being used in the research, storing the name and address of participants can be considered. This particular case focused on whether people were comfortable wearing all these devices, and didn't involve research on the data. Therefore there is no reason to link the data back to the data subject. However, if these methods are used in the future in social science, it might be useful to inform the participants of the results of the research or give them personal recommendations based on their data. If the researchers don't find a use for these data, they should not be stored to mimimise data. Data minimisation is one of the core aspects of Data Protection by Design [32].

Chapter 9

Project Conclusion

In this document we have presented our recommendations for complying with the GDPR regulations in research projects. Important to note is that these techniques can only be seen in the context of a certain research project. Therefore we have provided examples of implementations in our individual theses and a case study in the project discussion. There are multiple ways to implement the regulations and we do not claim that our recommendations are the only solution to the problem. Data officers should perform a risk analysis and given these risks multiple solutions are possible.

9.1 Future work

Due to time constraints we were not able to implement our recommendations in actual implementations. Given that the University of Groningen is building a system for storing research data, it would be interesting to see whether our ideas can help with realising this system.

It would also be interesting to re-evaluate this document when the GDPR is in effect in May 2018. It could be compared to other documents with technical implementations that are forthcoming, for example the currently missing chapters of the SDM. It would also be interesting to revisit this document when the first cases of violations of the GDPR occur, to see whether our implementations would have been sufficient to protect the privacy of the data subjects.

References

[1] Emily Steel, Callum Locke, Emily Cadman and Ben Freese, "How much is your personal data worth?", June 12, 2013, http://www.ft.com/cms/s/2/927ca86e-d29b-11e2-88ed-00144feab7de.html?ft_site=falcon

[2] European Union, "Directive 95/46/EC of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data", 24 October 1995, available at: http://www.refworld.org/docid/3ddcc1c74.html

[3] "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)", 27 April 2016, http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

[4] "The Standard Data Protection Model: A concept for inspection and consultation on the basis of unified protection goals". Die Bundesbeauftragte für den Datenschutz und die Informationsfreiheit, March 1 2017. https://www.datenschutzzentrum.de/uploads/SDM-Methodology_V1_EN1.pdf

[5] M. Hansen, M. Jensen and M. Rost, "Protection Goals for Privacy Engineering," 2015 IEEE Security and Privacy Workshops, San Jose, CA, 2015, pp. 159-166.

[6] "Top 10 Operational Impacts Of The GDPR: Part 2 - The Mandatory DPO". Iapp.org. N.p., 2017. Web. 22 May 2017. https://iapp.org/news/a/top-10-operational-impacts-of-the-gdpr-part-2-the-mandatory-dpo/

[7] "Top 10 operational impacts of the GDPR: Part 8 - Pseudonymization". Iapp.org. N.p., 2017. Web. 04 July 2017. https://iapp.org/news/a/top-10-operational-impacts-of-the-gdpr-part-8-pseudonymization/

[8] Kamara, I., "Co-regulation in EU personal data protection: the case of technical standards and the privacy by design standardisation 'mandate'", in European Journal of Law and Technology, Vol 8, No 1, 2017. http://ejlt.org/article/view/545/723#_ednref38

[9] Data Protection Directive. Web. 08 June 2017 http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:31995L0046

[10] "A Glossary of Terms and Definitions as used in relation to the GDPR.". Trunomi. Web. 11 June 2017 http://www.eugdpr.org/glossary-of-terms.html

[11] Article 29 Data Protection Working Party. Web. 07 June 2017 https://cnpd.public.lu/fr/publications/groupe-art29/wp216_en.pdf

[12] "Sensitive data and lawful processing", Bird & Bird lawfirm. Web. 07 June 2017 https://www.twobirds.com/~/media/pdfs/gdpr-pdfs/25--guide-to-the-gdpr--sensitive-data-and-lawful-processing.pdf?la=en

[13] "Why pseudonymization is not the silver bullet for GDPR." J. Ryan, Head of Ecosystem, PageFair. Web. 27 June 2017 https://pagefair.com/blog/2017/pseudonymization-gdpr/#_ftn2

[14] Y.-A. de Montjoye, C. Hidalgo, M. Verleysen and V. Blondel, "Unique in the Crowd: The privacy bounds of human mobility," Nature, no. 1376, 2013.

[15] "Why Pseudonyms Don't Anonymize: A Computational Re-identification Analysis of Genomic Data Privacy Protection Systems", Bradley Malin Data Privacy Laboratory, School of Computer Science. Carnegie Mellon University, Pittsburgh, Pennsylvania. Web. 24 June 2017 https://dataprivacylab.org/dataprivacy/projects/linkage/lidap-wp19.pdf

[16] "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule", U.S. Department of Health & Human Services. Web. 07 June 2017 https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

[17] L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000. Web. 07 June 2017 https://dataprivacylab.org/projects/identifiability/paper1.pdf

[18] "Handbook of Applied Cryptography", A. Menezes, P. van Oorschot, and S. Vanstone, CRC Press, 1996. Web. 15 June 2017. http://cacr.uwaterloo.ca/hac/about/chap1.pdf

[19] "The first collision for full SHA-1", Marc Stevens and Pierre Karpman, CWI Amsterdam in collaboration with Elie Bursztein, Ange Albertini and Yarik Markov, Google Research. Web. 25 June 2017 https://shattered.io/static/shattered.pdf

[20] "Salted Password Hashing - Doing it Right", Defuse Computer Security R & D, Web. 23 June 2017 https://crackstation.net/hashing-security.htm

[21] "Handbook of Applied Cryptography", A. Menezes, P. van Oorschot, and S. Vanstone, CRC Press, 1996. Chapter 8. Web. 15 June 2017. http://cacr.uwaterloo.ca/hac/about/chap8.pdf

[22] Securosis, *Understanding and Selecting a Tokenization Solution*, Web. 16 June 2017. https://securosis.com/assets/library/reports/Securosis_Understanding_Tokenization_V.1_.0_.pdf

[23] FOX-IT BV, *DigiNotar Certificate Authority breach "Operation Black Tulip"*, 5 September 2011, Web, 20 June 2017 https://www.rijksoverheid.nl/ministeries/ministerie-van-binnenlandse-zaken-en-koninkrijksrelaties/documenten/rapporten/2011/09/05/diginotar-public-report-version-1

[24] "Tokenization vs Encryption", Skyhigh Networks. Web. 24 June 2017 https://www.skyhighnetworks.com/cloud-security-university/tokenization-vs-encryption/

[25] "Polymorphic Encryption and Pseudonymisation", B. Jacobs and the PEP team, Radboud University Nijmegen. Slides from a talk on 22 February 2017 http://www.cs.ru.nl/B.Jacobs/TALKS/pep-overview.pdf

[26] "Polymorphic Encryption and Pseudonymisation for Personalised Healthcare", Eric Verheul, Bart Jacobs, Carlo Meijer, Mireille Hildebrandt, Joeri de Ruiter. Institute for Computing and Information Sciences Radboud University Nijmegen, The Netherlands. Web. 22 June 2017. https://eprint.iacr.org/2016/411.pdf

[27] "Anonymisation is great, but don't undervalue pseudonymisation", P. Lee, Partner, Privacy, Security and Information, Fieldfisher. Web. 27 June 2017 http://privacylawblog.fieldfisher.com/2014/anonymisation-is-great-but-dont-undervalue-pseudonymisation/

[28] Arvind Narayanan and Vitaly Shmatikov, Robust De-anonymization of Large Sparse Datasets, May 18 - 21, 2008, SP '08 Proceedings of the 2008 IEEE Symposium on Security and Privacy Pages 111-125, http://dl.acm.org/citation.cfm?id=1398064.

[29] Paul Ohm, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, August 13 2009, UCLA Law Review, Vol. 57, p. 1701, 2010; U of Colorado Law Legal Studies Research Paper No. 9-12., https://ssrn.com/abstract=1450006

[30] Nebeker C, Lagare T, Takemoto M, et al. Engaging research participants to inform the ethical conduct of mobile imaging, pervasive sensing, and location tracking research. Translational Behavioral Medicine. 2016;6(4):577-586. doi:10.1007/s13142-016-0426-4.

[31] "Chapter 5: Key definitions – Unlocking the EU General Data Protection Regulation", D. Gabel, T. Hickman, Web. 12 July 2017
`https://www.whitecase.com/publications/article/chapter-5-key-definitions-unlocking-eu-general-data-protection-regulation`

[32] "Art. 25 GDPR Data protection by design and by default", General Data Protection Regulation (GDPR). Web. 12 July 2017 `https://gdpr-info.eu/art-25-gdpr/`