D1.8 Public Availability of Research Data

WP1– Project Management

Version: 1.00

SPHINX

A Universal Cyber Security Toolkit for Health-Care Industry

Disclaimer

Copyright message

Document information

Grant Agreement Number	**826183**	**Acronym**	**SPHINX**
Full Title	A Universal Cyber Security Toolkit for Health-Care Industry		
Topic	SU-TDS-02-2018 Toolkit for assessing and reducing cyber risks in hospitals and care centres to protect privacy/data/infrastructures		
Funding scheme	RIA - Research and Innovation action		
Start Date	1st January 2019	**Duration**	36 months
Project URL	http://sphinx-project.eu/		
EU Project Officer	Reza RAZAVI (CNECT/H/03)		
Project Coordinator	Dimitris Askounis, National Technical University of Athens - NTUA		
Deliverable	D1.8. Public Availability of Research Data		
Work Package	WP1 – Project Management		
Date of Delivery	**Contractual**	M9	**Actual**	M9
Nature	R - Report	**Dissemination Level**	P - Public
Lead Beneficiary	NTUA		
Responsible Author	Christos Ntanos	**Email**	cntanos@epu.ntua.gr
		Phone	
Reviewer(s):	Sergiu Marin [Polaris Medical], Jason Mansell [TECNALIA]		
Keywords	Access Policies, Metadata, Long-Term Storage		

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

Document History

Version	Issue Date	Stage	Changes	Contributor
0.10	19/07/2019	Draft	ToC	George Doukas (NTUA)
0.20	29/08/2019	Draft	Contributions to Sections 2,3,4&5	George Doukas (NTUA)
0.30	03/09/2019	Draft	First Draft	George Doukas (NTUA)
0.40	03/09/2019	Draft	Final draft submitted for review	George Doukas (NTUA)
0.50	25/09/2019	Draft	Review 1	Sergiu Marin (POLARIS MEDICAL)
0.51	25/09/2019	Draft	Review 2	Jason Mansell (TECNALIA)
0.60	26/09/2019	Pre-Final	Incorporated all review comments	George Doukas (NTUA)
0.61	26/09/2019	Pre-Final	Quality Control	Michael Kontoulis(NTUA)
1.00	26/09/2019	Final	Final	Christos Ntanos (NTUA)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

Executive Summary

Sphinx will not have to store/access personal data for research purpose but it will use data created by the users within the three pilots of the platform. Even though there is no need for any personal information Sphinx has to employ the necessary techniques, to ensure that data privacy is respected. Since the adoption of the General Data Protection Regulation (GDPR) by all member states of the European Union (EU), all parties that handle and store personal data, referred to as data controllers, and parties that perform processing of said data, referred to as data processors, need to conform to specific guidelines that govern both the storage and processing of data and the ability and rights of data subjects to control their stored data. The present document covers the main techniques that will be employed by Sphinx to ensure private data protection and GDPR compliance.

At first, a review of the most commonly-used encryption and anonymisation techniques is given. Storing and processing data requires that certain guidelines of the GDPR are respected; these guidelines are briefly reviewed and the way Sphinx conforms to these guidelines is outlined.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

Table of Figures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

1 Introduction

1.1 Purpose & Scope

Conforming to the established legislature and respecting stakeholder's and physical person's sensitive information is a major requirement and challenge for every process, framework or application that deals with storage or processing of personal data. With the advent and adoption of the GDPR by all member states of the EU, the set of requirements, that each party that deals with personal data must fulfil, has been defined in a much stricter sense than before. Moreover, data regulation under the GDPR has been extended to not only cover the kind of data stored, but also explicitly define the rights of a physical person whose personal data are stored, commonly referred to as data subject under the GDPR terminology, *after* her/his consent is given and her/his data are stored. Under the present regulation, a data subject can revoke any consent given at any time and request that any stored personal data be removed. In general, data subjects now have the right to be aware of any details about how their data are processed, analysed, shared or used in decision-making and analytics and have moreover the right to access and/or erase these data.

Sphinx is not going to use within its research activities personal and sensitive data. Since the collection of mostly personal data within Sphinx coincides with the same data collection in the day-to-day operations of the pilot organisations, the consortium as a whole will be permitted access to aggregated and anonymised data from the repositories of the pilot partners for research and dissemination purposes.

Appropriate technical and organisational measures will be included, such as anonymisation / pseudonymisation and encryption of personal data and/or regularly testing and evaluating the effectiveness of the measures to ensure the data processing is secure. Even though data security burdens the processor and/or controller (in SPHINX's case the healthcare providers), SPHINX aims to offer a holistic solution that will protect the Health IT ecosystem against cyber-attacks, including data breaches. In other words, the SPHINX project aims to offer a security by design approach that will protect, among others, data subjects, namely patients, against possible cyber threats that would lead to violation of their personal data.

1.2 Structure of the deliverable

This deliverable documents the ways in which Sphinx will protect sensitive data, if any, and how it conforms to the GDPR. Section 2 provides briefly an overview of project's data management approach. Section 3 presents the main techniques currently used for anonymising and encrypting data, while also describing the ways that these techniques can be used to fully conform to the GDPR. Section 4 gives an overview regarding ethical aspects from the use of research data. Section 5 finally, provides the main conclusions.

1.3 Relation to other WPs & Tasks

The present deliverable is released within the context of Work Package 1 "Project Management". Deliverables D1.3 to D1.8 & D1.10 provide information about the type and the exploitation of collected data and the GDPR framework. In these deliverables an analytical description of how the Sphinx project complies with GDPR regulations, the processing activities of personal data that might take place in the Sphinx platform, and how the use of state of the art technologies will protect the rights of end-users in general.

1.4 List of Abbreviations

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

GDPR	General Data Protection Regulation
FAIR	Findability, Accessibility, Interoperability and Reusability
DMP	Data Management Plan

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

2 Data Summary

The Data Management Plan for the Sphinx project has been presented in deliverable D1.3 Data Management Plan, since Sphinx aims to participate in the Open Research Data Pilot in Horizon 2020. According to the guidelines of the Open Research Data Pilot in Horizon 2020[1], a detailed process must be described in D1.3 in order to identify the data to be collected, processed and/ or generated, formulate the methodology and standards to be applied, the data sharing, curation and preservation process to be adopted as well as the data curation and preservation mechanisms.

According to D1.3, a summary of the data management strategy is provided below:

- **Data Source and Acquisition**: data collected by Sphinx are both from its components' databases, but also from other relevant organisations, open data repositories and any other open data source.
- **Exploitation, availability of data, access, sharing, and re-use:** the project will ensure compliance with the FAIR data principles and will consider adopting pseudonymisation to personal data, if any, before sharing, accompanied by data encryption techniques.
- **Archiving, preservation and data security:** all data collected during the project will be stored in the projects database repository so that it can be compliant with the "right-to-be-forgotten" requests.

2.1 Open Access approach

Sphinx consortium has agreed to follow an "open access" approach (as much as possible depending on the specific data type) following the respective Horizon 2020 guidelines to ensure that the results of the project results provide the greatest impact possible. Sphinx will ensure the open access[2] to all peer-reviewed scientific publications relating to its results and will provide access to the research data needed to validate the results presented in deposited scientific publications. Publications and research data made available to third parties will not contain any personal information.

2.2 Roles related to the management of data

The current chapter identifies the Sphinx roles related to the management of data and their respective responsibilities, i.e. the roles of the data controller, data producer and data manager.

According to Article 4 of the EU GDPR, the Data Controller, the Data Processor and the Data Recipient are defined as follows:

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".

Processor – "means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller".

Recipient – "means a natural or legal person, public authority, agency or another body, to which the personal data are disclosed, whether a third party or not".

Data subject – "means a natural person whose personal data are processed and may be subject to pseudonymisation".

Third party – "means any entity other than the data subject, controller or processor"

[1] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

[2] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data- management/open-access_en.htm

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

Each and every organisation in the Sphinx project is solely responsible for their own data they create and upload. These data are not accessible to other organisations or individuals without prior permission. The decision of what data to store, how to process them and to whom they will permit access and under what conditions, befalls that organisation. Since the collection of mostly personal data within Sphinx coincides with the same data collection in the day-to-day operations of the pilot organisations, the responsibility cannot and should not be transferred to other parties.

Within this framework, the consortium as a whole will be permitted access to aggregated and anonymised data from the repositories of the pilot partners for research and dissemination purposes. The data that will be allowed to be used, the type and the extent of the anonymisation and aggregation algorithms that will be used will need to be approved by the organisations that will provide the data. A final approval for the dissemination of the aggregated and anonymised data from the pilot operation of the platform and for any use will be provided by the Innovation/Scientific Manager and the Technical Manager of the project.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

3 Data Security

The principle of data security requires that appropriate technical or organisational measures are implemented when processing personal data to protect the data against accidental, unauthorised or unlawful access, use, modification, disclosure, loss, destruction or damage. [3]The GDPR states that the controller and the processor should take into account "the state of the art, the costs of implementation and the nature, scope, context and purpose of processing, as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons" when implementing such measures. [4] Depending on the specific circumstances of each case, appropriate technical and organisational measures could include, for example, pseudonymising and encrypting personal data and/or regularly testing and evaluating the effectiveness of the measures to ensure the data processing is secure. The appropriateness of security measures must be determined on a case-by-case basis and reviewed regularly.

Sphinx platform will provide appropriate and sufficient security mechanisms and management procedures in order to ensure personal data, if any, protection through a clear process of data collection, pseudonymisation, and guarantee data integrity and reliability, ensuring the system's high performance operation. Thus, the system will have high levels of security and visibility regarding data transactions and other processes. Moreover, it needs to be mentioned that the issue of data security will also be updated and made more specific and holistic during the implementation of all the components of the project as more technical details on the platform's databases and data processing activities will be specified.

Sphinx will take all necessary measures in order to assure data security and mitigate any risks associated with storage and transmission of data. Given that the project is still in its infancy at the time of writing the current report, it is not possible to establish a fixed protocol for data security. Nevertheless, a list of measures is provided that the consortium can adopt according to the advancements and the detailed definition of the technological aspects and pilot scenarios and requirements.

The project will preserve the data collected and generated during the lifetime of the project in order to allow the achievement of the project´s vision and goals. Most of the research datasets will be shared as open data using the Linked Open Data standards and open licences according to the Open Research Data Pilot in Horizon 2020 in which the project participates. More details on the project's data security measures are detailed in D1.3.

3.1 Landscape analysis in privacy and anonymisation

In the present section, an overview of the main existing techniques for anonymising and encrypting data is given. Additionally, approaches on the identification of the users are also provided. The extent to which these techniques can be used to conform to GDPR requirements will also be described.

3.1.1 Background

When data are communicated between two parties, it is often a requirement that the contents of the communication remain secret to non-participating parties. Encryption techniques that tried to provide the required security were used since antiquity, usually to protect military secrets. Nowadays, personal sensitive data are also a main focus for data encryption; whenever two parties need to exchange information in such a way so that the contents remain hidden from eavesdroppers, an encryption scheme must be applied. Furthermore, when a party communicates with another party, each party needs to be able to verify that the other party is who it claims to be and not a pretender. The techniques used for both encryption and identity

[3] General Data Protection Regulation, Recital 39 and Art. 5 (1) (f); Modernised Convention 108, Art. 7

[4] General Data Protection Regulation, Art. 32 (1).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

verification (via the so-called digital signature), typically involve the employment of similar techniques, which fall under the domain of cryptography.

Apart from encryption, there are cases where the data contain both informational content that needs to be communicated publicly and personal data that need to remain inaccessible. There is no reason to encrypt for example the measurements and results that take place in a medical study, as these need to be accessible by all interested researchers. The personal data of the patients involved however, should not be disclosed. In these cases, techniques that perform data masking on the original data set need to be applied. The main difference between encryption and data masking is that a masked data set is in plain format and can be read by anyone but cannot be reversed to the original data in any way. In data masking, data privacy protection is ensured storing or transmitting masked data, without disclosing the transformations performed.

Before proceeding with an analysis of the existing techniques, the definition of the terms pseudonymisation and anonymisation is going to be provided as these are used in most legal documents and especially in the GDPR.

- **Pseudonymisation** means the transformation of data in such a way so that personal data cannot be retrieved without the usage of additional identifiers, not present in the transformed set.
- **Anonymisation** means the transformation of data in such a way so that personal data cannot be retrieved in any way from the transformed set.

The definition covers all types of data transformation so, technically, encryption can be used both for pseudonymisation (by keeping the decryption key in a separate database) and for anonymisation (by dropping the key or using one-way hashes). However, in the present section, we will cover encryption separately and we will focus on data masking techniques only when discussing pseudonymisation and anonymisation.

3.1.2 Techniques & methodologies

3.1.2.1 Data pseudonymisation and anonymisation

It is often the case that data that concern individuals need to be communicated publicly. Examples of such cases are medical studies, poll results and etc. While personal information is inevitably disclosed in these cases, it is necessary to restrict the amount of data published or stored to the bare minimum needed for each case. When personal identification is not needed, the data disclosed should be such that no identification can be performed by reverse engineering, ensuring effectively that the people described by data remain anonymous.

Data anonymisation can be performed by a variety of techniques, which are not needed to be performed exclusively [1]. The first step in anonymising the data set is performing removal or encryption of personal identifiable information (PII) [2]; these include information like name, address, id number etc. The mapping to the original data can be maintained in a separate database in case of pseudonymisation or be entirely discarded for anonymisation.

Removal of PII often is not enough for ensuring privacy however, since combinations of other information can still lead to the identification of a person, especially if said combinations occur rarely in the original data set. Combinations of *Personal Characteristics* data (also called Quasi-identifying attributes), like ethnicity and sex, are typically such combinations. If, for example, a single combination of a certain nationality, sex and marital status appears in a data set, this certain person can be identified by only requiring the extra knowledge that she/he is a member of the data set. Anonymising a set that contains personal characteristics data typically involves one or more of the techniques of **data masking**, namely a) encryption, b) shuffling, c) substitution, d) variance, e) masking and f) pruning.

Encryption, as already mentioned, is covered in a different section.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

Shuffling is the technique of randomly or via an algorithm, changing the values of a data column. The transformed data contain entries that cannot lead to the person's identification since the information contained in each row are now invalid. The transformed data sets may also, depending on the case, keep its main statistical characteristics and still be useful for statistical analysis. Shuffling has the drawback that when performed as the single anonymisation technique, can leave the transformed data set open to reverse engineering attempts, especially when the shuffling algorithm is known.

Substitution is similar to shuffling, the main difference being that the substituted values do not originate from values occurring in other entries of the data set, but rather in external lists. A list of all the ethnicities for example may be used to randomly change the ethnicity of an entry. In contrast with shuffling, this may lead to the appearance of values not appearing in the original data set.

Variance is typically applied to values of numerical nature and consist of adding or subtracting a random noise to the value; an example is to compute a random number between -5% and 5% of a person's height and add/subtract this to the original value. The variance technique can too lead to the formation of data sets that keep their original statistical characteristics, if the mean value of the random noise added is zero.

Masking is the substitution of all, or part of, the value of a field with null values or a standard character. It is typically used in credit card numbers where the unmasked fields can be used to identify details like type of card (VISA or MasterCard), but numbers unique to the individual person are being left out. Masking also applies to non-text data, like images or video. Substituting the face of minors in media pictures is a typical case of image masking.

Pruning is simply the deletion of a record. Pruning can be used in records that have rare combinations and are thus easier to be traced back to a certain person. As these combinations are rare, the statistical properties of the data will not be affected much, however there is the drawback that such combinations often have scientific value. A special case of pruning is the deletion of only certain values of a record by removing columns or assigning null values; this last technique is also called nulling. Nulling is efficient for discarding personal data that are not needed for processing but has the drawback that it can draw suspicion that data masking has been performed on the data set.

Except from Masking, all anonymisation techniques can be used in such a way that the transformed data set, though technically contains false data, appears realistic to an outside viewer.

Figure 1 depicts an example of data masking, with each column being affected by the transformation being denoted by the purple outline. Except from the masked credit card number and the null ages, all other values appear realistic. However, the combinations of name and sex lead to a male Helen and a female Jack which do not appear realistic. Data masking algorithms can be designed such that realistic looking combinations and values are produced in the transformed data set. The set of all the transformations performed defines a transformation matrix; the inverse of this transformation matrix may be used to restore the original data. When the transformation matrix is stored in a separate database, the data set is considered pseudonymised, whereas, if it is deleted, it is considered as anonymised.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

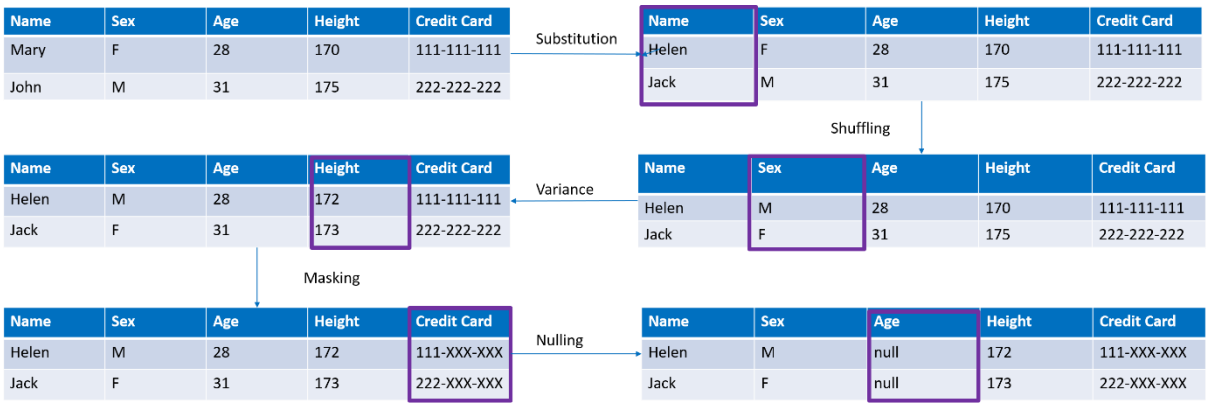

Figure 1: Example of data masking

3.1.2.1.1 Data anonymisation models for dealing with attackers having access to multiple data sets

Though in the previous sections we analysed the common masking techniques, the degree of anonymisation often depends not only on the efficiency of the transformation themselves, but to the extent that an attacker possesses knowledge of extra datasets containing information of the data subjects. These datasets may be also anonymised, but when combined they could be used to de-identify the subject(s); a typical example involves combination with clinical data released as anonymous with a public voter dataset [5]. To handle these cases, a set of privacy models have been suggested; each one attempts to quantify the degree of anonymisation to one or more metrics and transform the data set in a way that target values of the metrics are achieved.

In the following sub-sections a brief overview of the most commonly used models will be given.

3.1.2.1.1.1 k-anonymity

A data set is k-anonymous if each entry from the released data set cannot be distinguished from at least k-1 other entries. The distinction is based on the number of attributes identified as personal characteristics (also called quasi-identifiers). Each group that shares the same values of the quasi-identifiers is called an equivalence class. Figure 2 depicts an example of a k-anonymous set with k=2. Here, a set of four attributes have been designated as quasi-identifiers, namely the race, birth year, gender and postal code (which is masked).

To achieve k-anonymity two techniques are commonly being employed:

- Suppression is similar to masking and assigns a generic single value to a variable. This can be like the masked postal code of the example or a fake value if a realistic look of the data is needed
- Generalisation, in which groups of values are merged into one. For example, birthdates can be grouped into decades so that any birthdate from 1980 to 1989 being assigned the value of 1980.

Race	Birth	Gender	Postal Code	Diabetes
White	1980	male	20*	Yes
White	1980	male	20*	No
White	1980	male	20*	Yes
White	1982	female	18*	Yes
White	1982	female	18*	Yes
Black	1982	male	18*	Yes
Black	1982	male	18*	No

Figure 2: Example of a k-anonymous set with k=2 and three equivalent classes

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

One of the main problems of k-anonymity is that efficient generalisation depends on the proximity of values. If values are very dispersed, broad categories should be used to achieve a high value of *k*, and this leads to degradation of data (the data loses its usefulness to the recipients instead of only to the attackers).

Another issue is that what is considered a non-quasi-identifier by the holder of the data may be effectively be a quasi-identifier in the hands of an attacker. In the example given, diabetes is not part of the quasi-identifier set and was not transformed under k-anonymisation. However, attributes not transformed may exhibit a lack of diversity; if this is the case the attacker may use this to expose sensitive information. In the example given, the attacker may know that "Alice" is white and was born in 1982. Though the attacker cannot distinguish which of the two rows (4 or 5) correspond to Alice, she/he will know that Alice has diabetes. The problem can be remedied by considering *Diabetes* to be quasi-identifier too, however this tends to aggravate the problem of efficient generalisation described above.

As a final note, we should also mention an extension of *k-anonymity* model, namely the *k-map*. Similar to *k-anonymity*, the *k-map* requires that the entries cannot be distinguished from at least *k-1* entries, in the case of *k-map* however the records counted correspond to the larger population and only to those present in the data set and thus the computed k values are adjusted accordingly along with the characteristics of the relevant transformations. Consider for example the postal code as a quasi-identifier and that we perform k-anonymity with *k=5*. Suppose further that a postal code appears in the dataset that corresponds to a town with 20 registered residents. An attacker can identify this, and reduce her/his search space greatly. To remedy this, we can mask the postal code. While this transformation may not alter the k-value relative to the original data set, it will greatly increase it relevant to the larger population

3.1.2.1.1.2 l-diversity

With l-diversity only equivalent classes that have a certain amount of diversity in their sensitive data are contained in the data set. The diversity is measured by the parameter *l* and there are various ways to define the measure of "l-diversity" [6]:

- distinct: where each equivalent class is required to have at least *l* distinct values in their sensitive data
- entropy: where the entropy of each equivalent class *E* denoted by *Entropy(E),* is greater or equal to the logarithm of *l. Entropy(E)* is defined as: $Entropy(E) = -\sum_s p(E,s)\log\left(p(E,s)\right)$ with p(E,s) being the fraction of records in E that have the sensitive value s.
- (c-l) recursive: where the counts $n(E,s_i)$ are computed for each equivalent class and each distinct sensitive value s_i (so that $n(E,s_1)$ is the number of times that s_1 appears in *E*, $n(E,s_2)$ is the number of times that s_2 appears in *E*, and so on). The counts are then sorted in descending order and a constant *c* is chosen. If n_{max} is the first element of the list of sorted counts, (c-l) reclusiveness requires that: $n_{max} \leq c\sum_{n\neq n_{max}} n$

For the example of Figure 2, if distinct diversity is chosen, the data set has a diversity equal to one.

Though l-diversity achieves better results in terms of anonymisation than simple k-anonymity, it too can have its limitation depending on the characteristics of the data set. If a sensitive value is very rare (e.g. a rare disease), it is difficult to achieve a high value of *l*. Furthermore, the equivalent classes may be low when compared to the count of the data entries to achieve a satisfactory value of *l*. For 10000 entries for example, we must have a maximum of 100 equivalent classes to achieve an *l* value equal to 2. As a final note, *l-diversity* does not distinguish between the semantics of sensitive values. This has as a result that entries are considered diverse although semantically they are not. Consider the example depicted in Figure 3. Although the equivalent class has a diversity equal to 2, an attacker with partial information may infer that "Alice" has a serious medical condition.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

Race	Birth	Gender	Postal Code	Condition
White	1982	female	18*	HIV+
White	1982	female	18*	Breast Cancer

Figure 3: Example of diverse, yet semantically linked, sensitive information

3.1.2.1.1.3 t-closeness

Intuitively, an equivalent is defined to have t-closeness relative to a sensitive attribute, if the distribution that the sensitive attribute has in the equivalent class, is similar to the one it has in the whole data set. More precisely, if the distance between the distribution of the sensitive variable in the equivalent class and the distribution of the sensitive variable in the data set is at most *t*, then the equivalent class has t-closeness relative to the attribute. If all of the equivalent classes have t-closeness, then the whole data set is said to also have t-closeness.

t-closeness ensures a good level of anonymity on the data, however it has the common problem of all anonymisation techniques, mainly that what is considered sensitive data may in fact be a quasi-identifier in the hands of the attacker. For example, an attacker could, by external means, know that a person has some, yet unknown, medical condition and search the data set to identify the exact condition or at least narrow down the list.

3.1.2.1.1.4 δ-presence

For an equivalent class, δ_E is defined as the ratio between the size of an equivalent class that is present in a dataset and the size of the same equivalent class as this is considered in the general population. Consider as an extreme case that a town with postal code equal to 12345 has a total of five residents and that a dataset contains 4 subjects with the same postal code. Then we know that 4 out of the 5 who live in that town are part of the dataset. In numbers the δ_E parameter for this equivalent class has a value of 4/5, or 80%. In effect, δ_E is the ratio between the k-anonymous and k-map equivalent classes.

It is evident that for anonymised data sets, the δ_E for each equivalent class, should be the smallest possible. In essence, if we define the δ of the data set to be the maximum value of the δ values of each equivalent classes, then δ should be as low as possible.

Although δ is a great indicator for anonymisation, the main difficulty lies with estimating the data set that represents the larger population. This data set is, in many cases, not available so that estimations may be given by the data experts.

It is noteworthy that δ-presence, as originally was proposed [7], involved the computation of two δ parameters: δ_{max}, which is the same described above and δ_{min}, which was the *minimum* value of the δ values of each equivalent classes. This was to compensate for symmetry attacks, namely attacks that were based on the knowledge that someone was *not* in the case. However, this case is not encountered frequently in practice.

3.1.2.1.2 Data Anonymisation via aggregation

A special case of data anonymisation is when only aggregated data are needed and information corresponding to specific individuals is of no concern. If, for example, a survey needs to check for correlations between nutrition habits and specific health markers, once the correlations are obtained the original data in row format are no longer needed. It is, thus, possible to store only the aggregated data needed for correlations and discard the original data set; the aggregated data can be considered as anonymised.

One common technique for data aggregation is the Data Cube [3]. A Data Cube as an $m_1 x m_2 x ... x m_n$ array, with n being the number of aggregated variables and m_i being the number of distinct values the variable indexed by

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

i can take. Figure 4 depicts an example of a Data Cube for the case of *n=3*. If a certain combination has a very low count when compared to the average value, then we discern that this combination is a rare one and can lead to identification of the subject's details. Such combinations are thus dropped. For further security a random small noise with a mean value of zero can be added to each cell, so that the data set is further obfuscated.

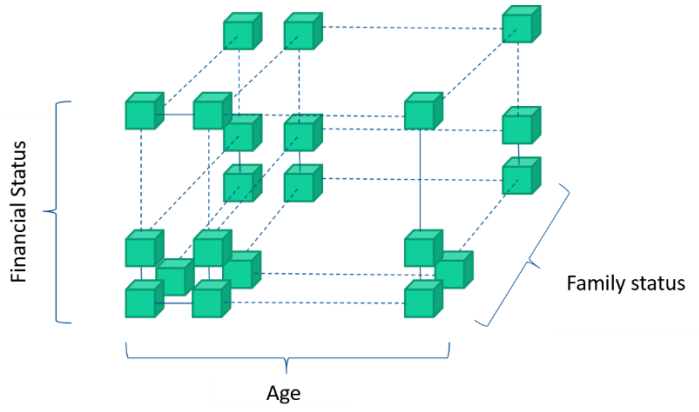

Figure 4: Data Cube Example

3.1.2.2 Pseudonyms

Broadly speaking, a pseudonym is a set of artificial identifiers that replace PII. In contrast with anonymisation of PII covered in the previous section, pseudonyms are mainly applicable to users of a service of platform and are typically used to cover the personal details of a platform user, as she/he appears on the platform. These personal details do not only cover the typical PII, like name, address etc., but also details of usage that can be linked back to the user. The subject's IP address for example, can be used to trace back her/his location and should therefore be hidden when using a pseudonym. Usage of pseudonym is required when there is a need that a subject appears under a false identity. The identity of an informant contacting the police for example, should remain secret to the public, however the police should be able to identify her/him.

Typically, pseudonyms are generated via an IP masking infrastructure such as TOR. Users negotiate a key to establish a secure connection with the pseudonym issuing authority; this authority is typically the platform that implements the pseudonym infrastructure. The authority issues a certificate for the pseudonym, which is then used by the user for authenticating her/his pseudo-identity. The authority keeps the correspondence between the personal identity and the pseudo-identity and appropriately authorises the pseudo-user and audits her/his activities to discern any attempt of pseudonym hijacking.

3.1.2.3 Encryption

Encryption of data refers to a set of techniques that can be used between two parties to exchange information in a secure and reliable way. This means that data exchanged between them are:

- Encrypted: No other party can make sense of the data unless the other party has possession of the private key needed to decipher the information.
- Signed: The identity of a sender can be verified in a way that the recipient is sure that the sender is who she/he claims to be. Upon receival of the message, the sender cannot deny that the message originated by

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

her/him and cannot claim that the contents of the message were others than those received by the receiver.

Apart from data exchange, encryption can also be used when storing data. There are two kinds of encryption, the symmetric/private key and the public key encryption. In symmetric key encryption, the key for both encryption and decryption is the same and is shared between communicating parties. Public key encryption on the other hand relies on the existence of two keys: one public and one private. The public key is used for encryption and, as its name implies, is publicly shared. Anyone who wishes to send an encrypted message to the receiver, does so by using the public key. The private key is kept at the receiver's side only and is inaccessible by the public. The receiver uses the private key to decrypt the message.

In general, public key encryption is used more widely. It has several benefits over symmetric encryption, most notably the fact that communicating does not require the exchange of keys beforehand, a process that can be both time-consuming and introduce additional risk. The general principles of generating and using public keys for cryptography are depicted in the next three figures.

Figure 5 depicts the process of generating a private/public key pair. The subject uses a key generation algorithm which generates a key pair. Key pair generation algorithms typically make use of large random numbers; the main idea behind most public key cryptography algorithms is that though multiplication of two large numbers is a computationally easy process, the reverse procedure, factoring a product of two large prime numbers, is an intractable process. After generating a key pair, this key pair can be used for secure communication and digital signing as depicted in Figure 6 and Figure 7 respectively. In the first case, anyone wishing to communicate with the receiver, uses the receiver's public key to encrypt the message. Upon receival of the encrypted message, also called a cipher, the receiver can use her/his private key to decrypt the message. In the digital signing case on the other hand, the sender uses her/his own private key to sign the message. The receiver, upon getting the signed document, uses the sender's public key to obtain the original document. If any modification was made while the message was en-route, the verification procedure will fail.

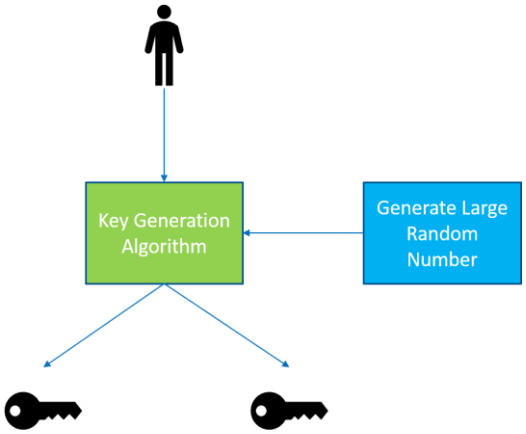

Figure 5: Public Private Key pair generation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

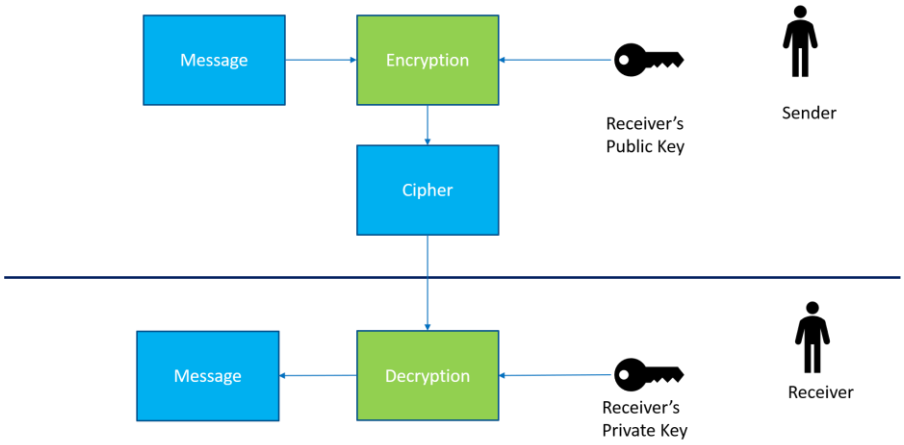

Figure 6: Encryption and Decryption of a message under the public key encryption scheme

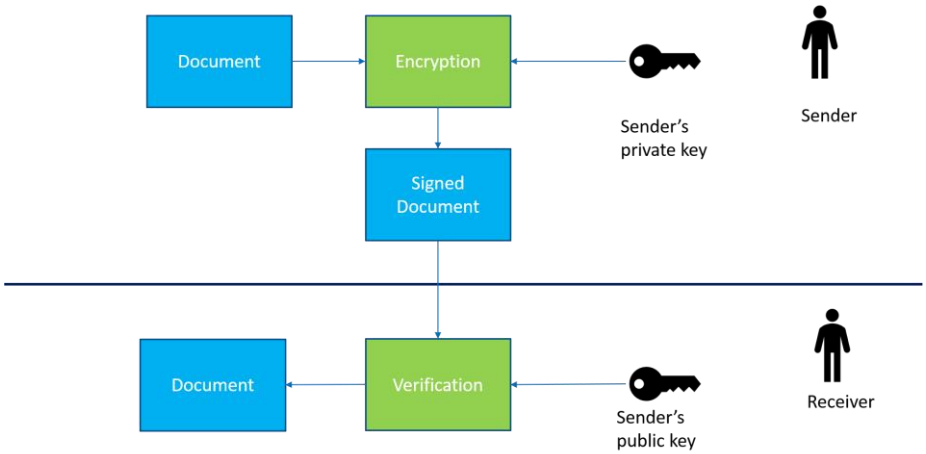

Figure 7: Digital signature under the public key encryption scheme

Although the main mechanisms of public cryptography are simple, when a large network of users are being connected and wish to exchange information securely, it can be quite difficult to distribute public keys in such a way such that the end users can be sure that the party publishing the public key is the actual owner of the public key. In order to distribute public keys securely, networks are usually called to implement what is commonly known as a Public Key Infrastructure (PKI). A PKI can create, distribute and revoke digital certificates, which are digital documents that prove the ownership of a public key.

The establishment of a binding between an entity and a public key is verified by the Certificate Authority (CA). The CA issues certificates, provided that registration was carried through correctly. The confirmation of correct registration is done by the Registration Authority (RA), which in many cases can be the same organisation as the CA. A third entity, called the Voucher Authority (VA), can vouch on behalf of the CA for the validity of the entity's information. Though a VA can reject a party's request if validation fails, it cannot issue or revoke the certificate; this is solely the responsibility of the CA.

Once the communicating parties can share public keys in a trustworthy manner, a security protocol can be used to encrypt communication. The Transport Layer Security (TLS) and the soon to be deprecated Secure Sockets Layer (SSL) are the most common examples of such protocols. Under TSL, before initiating exchange of data, a handshake between parties must first take place. In the handshake at least one party (the server) provides its certificate thus proving its identity. After validation, both parties generate and exchange keys and encrypt their messages using a private key algorithm. Figure 8 depicts an example of a PKI setup.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

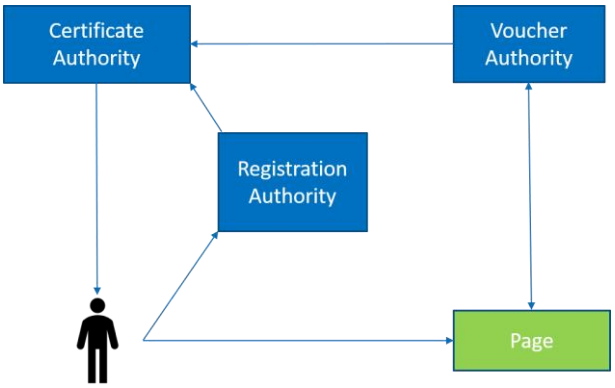

Figure 8: Typical PKI Infrastructure

3.2 GDPR Compliance

Since the adoption of GDPR by all member states of the EU, the processes of storing, handling and transmitting data need to conform to specific rules in order to be considered GDPR compliant. Under the GDPR any entity that handles personal data is fully accountable for conforming to the GDPR and responsible for any data breaches.

Entities that store and process personal data are referred to in the GDPR as *data controllers*. The exact definition as this appears in Article 4 of the GDPR reads: *"'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; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law;"*

An entity that does not collect personal data itself, but processes it on behalf of a controller is termed under the GDPR as *processor.* The exact definition of a processor as this appears in the GDPR is given as: *"'processor' means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller;"*

While listing the complete set of rules of GDPR is outside the scope of the present deliverable, controllers and processors should follow a set of guidelines.

These guidelines are summarised as follows:

- Legitimate interest: To store data that contain personal information, the controller should have a legitimate and valid reason to do so.
- Provision of Consent: Storing personal information of an individual requires that the said individual will give her/his consent.
- Right to be forgotten: An individual cannot only revoke his consent at any time, but she/he may request that any personal information already stored be removed upon request. In case of erroneous data, the individual may request correction of data.
- Right of access: Any processing of an individual's data must be approved by her/him.
- Data breaches: When any compromise of the system leads to personal data exposure, the supervisory authority should be notified promptly within 72 hours both of the occurrence of the breach and of the details concerning which data have been exposed.
- Anonymisation: GDPR encourages controllers to perform pseudonymisation on data before they are stored.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

- Sharing by third parties: Explicit consent must be given before personal data can be shared with third parties. The consent may be revoked at any time.

In the context of Sphinx platform there is no need for storing personalised information. As already mentioned, the consortium as a whole will be permitted access to aggregated and anonymised data from the repositories of the pilot partners for research and dissemination purposes. Sharing with third parties will also not take place in Sphinx.

3.2.1 Anonymisation/pseudonymisation techniques

Since the adoption of GDPR by all member states of the EU, the processes of storing, handling and transmitting data need to conform to specific rules in order to be considered GDPR compliant. Under the GDPR any entity that handles personal data is fully accountable for conforming to the GDPR and responsible for any data breaches.

According to the principle of storage limitation contained in both the GDPR and Modernised Convention 108, data must be kept "in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed".[5] Consequently, data would have to be erased or anonymised if a controller wanted to store them after they were no longer needed and no longer served their initial purpose. The process of anonymising data means that all identifying elements are eliminated from a set of personal data so that the data subject is no longer identifiable.[6] In its Opinion 05/2014, the Article 29 Working Party analyses the effectiveness and limits of different anonymisation techniques.[7] It acknowledges the potential value of such techniques, but underlines that certain techniques do not necessarily work in all cases. To find the optimal solution in a given situation, the appropriate process of anonymisation should be decided on a case-by-case basis. Irrespective of the technique used, identification must be prevented, irreversibly. This means that for data to be anonymised, no element may be left in the information which could, by exercising reasonable effort, serve to re-identify the person(s) concerned.[8]The risk of re-identification can be assessed by taking into account "the time, effort or resources needed in light of the nature of the data, the context of their use, the available re-identification technologies and related costs".[9] When data have been successfully anonymised, they are no longer personal data and data protection legislation no longer applies. The GDPR provides that the person or organisation controlling the personal data processing cannot be obliged to maintain, acquire or process additional information to identify the data subject for the sole purpose of complying with the regulation. However, this rule has a significant exemption: whenever the data subject, for the purpose of exercising the rights of access, rectification, erasure, restriction of the processing and data portability, provides additional information to the controller enabling his or her identification, then those data which were previously anonymised become personal data again.[10]

ISO/TS 25237:2017 defines anonymisation as a 'process by which personal data is irreversibly altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party' [ISO, 2017]. Similarly, NIST refers to anonymisation[11] as a 'process that removes the association between the identifying dataset and the data subject in simple words, an anonymised

[5] Ibid., Art. 5 (1) (e); Modernised Convention 108, Art. 5 (4) (e)

[6] General Data Protection Regulation, Recital 26.

[7] Article 29 Working Party (2014), Opinion 05/2014 on Anonymization Techniques, WP216, 10 April 2014

[8] General Data Protection Regulation, Recital 26.

[9] Council of Europe, Committee of Convention 108 (2017), Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data, 23 January 2017, para. 6.2.

[10] General Data Protection Regulation, Art. 11.

[11] (NIST) https://csrc.nist.gov/glossary/term/anonymization

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

dataset does not allow for identifying any individual, for either the controller or third party. Therefore, anonymised data do not qualify as personal data.

Personal information contains attributes, such as name, date of birth, sex, address, or other elements that could lead to identification. The process of pseudonymising personal data means that these attributes are replaced by a pseudonym.

EU law defines 'pseudonymisation' as "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".[12] Contrary to anonymised data, pseudonymised data are still personal data and are therefore subject to data protection legislation. Although pseudonymisation can reduce security risks to the data subjects, it is not exempt from the scope of the GDPR. The GDPR recognises various uses of pseudonymisation as an appropriate technical measure for enhancing data protection, and is specifically mentioned for the design and security of its data processing.[13] It is also an appropriate safeguard that could be used to process personal data for purposes other than for which they were initially collected.[14] Pseudonymisation is not explicitly mentioned in the legal definition of the CoE Modernised Convention 108. However, the Explanatory Report of Modernised Convention 108 clearly states that "the use of a pseudonym or of any digital identifier/ digital identity does not lead to anonymisation of the data as the data subject can still be identifiable or individualised".[15] One way to pseudonymise data is through data encryption. Once data has been pseudonymised, the link to an identity exists in the form of the pseudonym plus a decryption key. Without such a key, it is difficult to identify pseudonymised data. However, for those entitled to use the decryption key, re-identification is easily possible. The use of encryption keys by unauthorised persons must be particularly guarded against.

In broad terms, pseudonymisation refers to the process of de-associating a data subject's identity from the personal data being processed for that data subject. Typically, such a process may be performed by replacing one or more personal identifiers, i.e. pieces of information that can allow identification (such as e.g. name, email address, social security number, etc.), relating to a data subject with the so-called pseudonyms, such as a randomly generated values.

To this end, the ISO/TS 25237:2017 standard defines pseudonymisation as a 'particular type of deidentification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms' [16] [ISO, 2017]. Deidentification, according to the same standard is a 'general term for any process of reducing the association between a set of identifying data and the data subject'. A pseudonym is also defined as 'a personal identifier that is different from the normally used personal identifier and is used with pseudonymized data to provide dataset coherence linking all the information about a data subject, without disclosing the real world person identity'. As a note to the latter definition, it is stated in ISO/TS 25237:2017 that pseudonyms are usually restricted to mean an identifier that does not allow the direct derivation of the normal personal

Encryption techniques

Encryption aims at ensuring – via appropriately utilizing mathematical techniques - that the whole dataset that is being encrypted, is unintelligible to anyone but specifically authorised users, who are allowed to reverse this unintelligibility (i.e. to decrypt). To this end, encryption is a main instrument to achieve confidentiality of

[12] Ibid., Art. 4 (5).

[13] Ibid., Art. 25 (1).

[14] Ibid., Art. 6 (4).

[15] Explanatory Report of Modernised Convention 108, para. 18.

[16] Similar definition of pseudonymisation has also been adopted by the US National Institute of Standards and Technlogy (NIST)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

personal data by hiding the whole dataset and making it unintelligible to any unauthorised party (as long as state-of-the-art algorithms and key lengths are used and the encryption key is appropriately protected).

Under Article 32 of GDPR, controllers are required to implement risk-based measures to protect data security. One such measure is the "encryption of personal data" that "renders the data unintelligible to any person who is not authorized to access it." Cryptographic primitives can in general be used in pseudonymisation techniques to generate pseudonyms with desired properties.

Appropriate combination of several cryptographic schemes may also provide robust pseudonymisation approaches, for example by the use of techniques such as secure multi-party computation and homomorphic encryption. Several advanced cryptography-based pseudonymisation solutions have been proposed to alleviate data protection issues, especially in cases of personal data processing that present very high risks.

Several other well-known techniques, such as masking, scrambling and blurring, can also be considered in the context of pseudonymisation, having though restrictions with regard to their possible applications

Despite the different terminology used, in all the aforementioned definitions, it is clear that Sphinx focuses on using anonymised datasets.

If for some reason this must be altered the pseudonymisation process will focus on providing a high level of protection of data, providing access only to the pseudonymised data so as to render pseudonymisation a realistic choice. To this end, data controllers in Sphinx should have a clear understanding of the scope of data pseudonymisation and select the appropriate technique that could suffice for this particular scope.

SPHINX project aims to offer a security by design approach that will protect, among others, data subjects, namely patients, against possible cyber threats that would lead to violation of their personal data. In line with this, two components have been included in Sphinx architecture; namely the Anonymisation and Privacy (AP) component and Homomorphic Encryption (HE) component. The anonymisation module of AP component is a dataflow tool that offers user-defined transformations to clean, bake, structure, anonymise and or even encrypt data, while HE component implements the Homomorphic Encryption technique to ensure user data privacy and security.

3.2.2 Informed Consent

The informed consent form, which each participant will be asked to complete prior to their participation in the pilots, aims at ensuring that the user accepts participation and is informed about all relevant aspects of the research project; it will be collected in written form after the users have been provided with clear and understandable information about their role (including rights and duties), the objectives of the research, the methodology used, the duration of the research, the possibility to withdraw at any time, confidentiality and safety issues, risks and benefits. Moreover, pertaining to non-academic participants, an information sheet has also to be drafted that it will clearly outline the nature of the current research project.

The basic elements of the Sphinx consent form will include information about the following:

- Data collected
- Usage of users' data by third parties
- Users' rights concerning their data
- Explanation of why Sphinx processes user data
- Contact details

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

It is noted that the notion of consent and all matters related to it as far as personal data processing is concerned shall be examined in detail under Deliverable D1.10.

- Sphinx information sheet will include information about the following topics:
- Short introduction to Sphinx as well as its vision and main goal
- Organizations responsible for the data collected during the project
- Data gathering activities during the project's duration
- Risks involved by participating in the project
- User's rights as a participant

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

4 Summary and Conclusions

In the present document the technical details of encryption and anonymisation techniques that are going to be used in Sphinx were presented. The algorithms and modules to be implemented are in line with the technical requirements that derive from the need for GDPR conformity. As long as these procedures/modules will be used at any time, Sphinx will be GDPR compliant concerning research data privacy requirements.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.

Annex I: **References**

[1] U.S. Department of Health and Human Services Office for Civil Rights. HIPAA Administrative Simplification Regulation, 45 CFR Parts 160, 162, and 164. 2013

[2] Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. J Am Med Inform Assoc. 2010;17(2):169–77.

[3] Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals, Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, Hamid Pirahesh, January 1997, Data Mining and Knowledge Discovery: Volume 1 Issue 1, 1997

[4] Confident Ltd. "Anonymous Surfing, AnonIC.org, 2004

[5] Weaving technology and policy together to maintain confidentiality.

Sweeney, L. Journal of Law, Medicine and Ethics. 1997, 25:98-110. (impact factor 1.04) Cited and discussed in the commentary of the HIPAA Privacy Rule.

[6] Machanavajjhala, Ashwin; Kifer, Daniel; Gehrke, Johannes; Venkitasubramaniam, Muthuramakrishnan (March 2007). "L-diversity: Privacy Beyond K-anonymity". ACM Trans. Knowl. Discov. Data.

[7] Mehmet Ercan Nergiz, Maurizio Atzori, Chris Clifton: Hiding the presence of individuals from shared databases. SIGMOD Conference 2007: 665-676

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826183 - Digital Society, Trust & Cyber Security E-Health, Well-being and Ageing.