EUROPEAN UNION AGENCY
FOR CYBERSECURITY

DATA PROTECTION ENGINEERING

From Theory to Practice

JANUARY 2022

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, the European Union Agency for Cybersecurity contributes to EU cyber policy, enhances the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperates with Member States and EU bodies, and helps Europe prepare for the cyber challenges of tomorrow. Through knowledge sharing, capacity building and awareness raising, the Agency works together with its key stakeholders to strengthen trust in the connected economy, to boost resilience of the Union's infrastructure, and, ultimately, to keep Europe's society and citizens digitally secure. More information about ENISA and its work can be found here: www.enisa.europa.eu.

CONTACT

For contacting the authors please use isdp@enisa.europa.eu
For media enquiries about this paper, please use press@enisa.europa.eu

CONTRIBUTORS

Claude Castelluccia (INRIA),
Giuseppe D'Acquisto (Garante per la Protezione dei Dati Personali),
Marit Hansen (ULD),
Cedric Lauradoux (INRIA),
Meiko Jensen (Kiel University of Applied Science),
Jacek Orzeł (SGH Warsaw School of Economics)
Prokopios Drogkaris (European Union Agency for Cybersecurity).

EDITORS

Prokopios Drogkaris (European Union Agency for Cybersecurity),
Monika Adamczyk (European Union Agency for Cybersecurity).

ACKNOWLEDGEMENTS

We would like to thank the colleagues from the European Data Protection Board (EDPB), Technology Subgroup and the colleagues from the European Data Protection Supervisor (EDPS), Technology and Privacy Unit, for reviewing this report and providing valuable comments.

We would also like to thank Kim Wuyts, Veronica Jarnskjold Buer, Konstantinos Limniotis, Paolo Balboni, Stefan Schiffner, Jose M. del Alamo, Irene Kamara and the ENISA colleague Athena Bourka for their review and valuable comments.

LEGAL NOTICE

COPYRIGHT NOTICE

ISBN 978-92-9204-556-2, DOI 10.2824/09079

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The evolution of technology has brought forward new techniques to share, process and store data. This has generated new models of data (including personal data) processing, but also introduced new threats and challenges. Some of the evolving privacy and data protection challenges associated with emerging technologies and applications include: lack of control and transparency, possible reusability of data, data inference and re-identification, profiling and automated decision making.

The implementation of the GDPR data protection principles in such contexts is challenging as they cannot be implemented in the traditional, “intuitive” way. Processing operations must be rethought, sometimes radically (similar to how radical the threats are), possibly with the definition of new actors and responsibilities, and with a prominent role for technology as an element of guarantee. Safeguards must be integrated into the processing with technical and organisational measures. From the technical side, the challenge is to translate these principles into tangible requirements and specifications by requirements by selecting, implementing and configuring appropriate technical and organizational measures and techniques

Data Protection Engineering can be perceived as part of data protection by Design and by Default. It aims to support the selection, deployment and configuration of appropriate technical and organizational measures in order to satisfy specific data protection principles. Undeniably it depends on the measure, the context and the application and eventually it contributes to the protection of data subjects’ rights and freedoms.

The current report took a broader look into data protection engineering with a view to support practitioners and organizations with practical implementation of technical aspects of data protection by design and by default. Towards this direction this report presents existing (security) technologies and techniques and discusses possible strengths and applicability in relation to meeting data protection principles as set out in Article 5 GDPR.

Based on the analysis provided in the report, the following conclusions and recommendations for relevant stakeholders are provided below:

Regulators (e.g. Data Protection Authorities and the European Data Protection Board) should discuss and promote good practices across the EU in relation to state-of-the-art solutions of relevant technologies and techniques. EU Institutions could promote such good practices by relevant publicly available documents.

The research community should continue exploring the deployment of (security) techniques and technologies that can support the practical implementation of data protection principles, with the support of the EU institutions in terms of policy guidance and research funding.

Regulators (e.g. Data Protection Authorities and the European Data Protection Board) and the European Commission should promote the establishment of relevant certification schemes, under Article 42 GDPR, to ensure proper engineering of data protection.

1. INTRODUCTION

Technological advancements over the last years have impacted the way our personal data is being shared and processed. The evolution of technology has brought forward new techniques to share, process and store data. This has generated new models of data (including personal data) processing, but also introduced new threats and difficulties for the end user to understand and control the processing. Continuous online presence of end users has resulted in an increased processing of large amounts of personal data at daily basis. Think of online shopping or using a mobile application to navigate to a specific location or contact friends and family. The whole data lifecycle has been augmented with many actors being involved and eventually end users not being able to fully understand and control who, for how long and for what purpose has access to their personal data.

These new technologies have often been introduced without a prior assessment of the impact on privacy and data protection. In this context, processing of personal data is often characterised by the absence of a predetermined purpose and by the discovery of new correlations between the observed phenomena, for example in the case of big data or machine learning. This modus operandi conflicts essentially with the principles of necessity and purpose limitation, as these are stipulated by the GDPR. Blockchain and distributed ledger technologies, as another example, offer the opportunity of replacing intermediation-based transactions, but at the potential expense of a substantial loss of individuals' control over their data, which remain visible in the chain by all blockchain participants, as long as it is active or perhaps even beyond that. This, depending of course on the use case, contradicts the GDPR principle of data minimization, and constitutes a severe obstacle for the exercise of the right to deletion by data subjects. Lastly, Artificial Intelligence systems might be empowered to take decisions with some degree of autonomy to achieve specific goals, for example in credit score evaluation in the finance domain. Such autonomy might very well be in conflict with the prerequisites of human agency over machines and self-determination, both at the heart of personal data protection and the GDPR.

The evolution of technology has brought forward new techniques to share, process and store data which has introduced new threats and challenges

As also discussed in [1], some of the evolving privacy and data protection challenges associated with emerging technologies and applications include: lack of control and transparency, incompatible reuse of data, data inference and re-identification, profiling and automated decision making. The implementation of the GDPR data protection principles in such contexts is challenging as they cannot be implemented in the traditional, "intuitive" way. Processing operations must be rethought and redesigned, sometimes radically (similar to how radical the threats and the attack vectors are), possibly with the definition of new actors and responsibilities, and with a prominent role for technology as an element of guarantee. Appropriate technical and organisational measures, as well as safeguards, must be considered at the earliest stage possible and integrated into the processing. This is the scope of the notion of data protection by design, enshrined in Article 25 of the GDPR.

1.1 DATA PROTECTION BY DESIGN

Data Protection by design has been a legal obligation since the GDPR came into effect in 2018. However, the concept emerged several years ago in the context of privacy engineering[1]. At that time named as Privacy by Design (PbD), it gained a lot of traction and was recognised as an essential component of practically implementing privacy and personal data protection. Nowadays, it is regarded as a multifaceted concept: in legal documents on one hand, it is generally described

[1] It was brought forward by Dr. Ann Cavoukian and was also evident as a notion but not explicitly mentioned in the Directive 95/46/EC and the ePrivacy Directive. See also European Data Protection Supervisor (EDPS) [6] EDPS Opinion 5/2018 "Preliminary Opinion on privacy by design"

in very broad terms as a general principle; by researchers and engineers on the other hand it is often equated with the use of specific Privacy Enhancing Technologies (PETS). However, privacy by design is neither just a list of principles nor can it be reduced to the implementation of specific technologies. In fact, it is a process involving various technological and organizational components, which implement privacy and data protection principles by properly and timely deploying technical and organization measures that include also PETS.

The obligation described in Article 25 is for controllers to have effective data protection designed and integrated into the processing of personal data, with appropriate default settings configuration or otherwise available throughout the processing lifecycle. Further to the adoption of the GDPR, the EDPB has published a set of guidelines [2] on Data Protection by Design and by Default and provided guidance on their application. The core obligation is the implementation of appropriate measures and necessary safeguards that provide effective implementation of the data protection principles and, consequentially, data subjects' rights and freedoms by design and by default. Through the various examples provided, it is evident that proper and timely development and integration of technical and organizational measures into the data processing activities play a big role in the practical implementation of different data protection principles.

Engineering those principles relates not only to choices made with regards to designing the processing operation but also selecting, deploying, configuring and maintaining appropriate technological measures and techniques. These techniques would support the fulfilment of the data protection principles and offer a level of protection adequate to the level of risk the personal data are exposed to. Data Protection Engineering can be perceived as part of data protection by Design and by Default. It aims to support the selection, deployment and configuration of appropriate technical and organizational measures in order to satisfy specific data protection principles. Undeniably it depends on the measure, the context and the application and eventually it contributes to the protection of data subjects' rights and freedoms.

Data Protection Engineering aims to support the selection, deployment and configuration of appropriate technical and organizational measures in order to satisfy data protection principles

1.2 SCOPE AND OBJECTIVES

The overall scope of this report is to take a broader look into data protection engineering with a view to support practitioners and organizations with practical implementation of technical aspects of data protection by design and by default. Towards this direction this report attempts to present existing (security) technologies and techniques and discuss possible strengths and applicability in relation to meeting data protection principles. This work is performed in the context of ENISA's tasks under the Cybersecurity Act (CSA)[2] to support Member States on specific cybersecurity aspects of Union policy and law relating to data protection and privacy. This work is intended to provide the basis for further and more specific analysis of the identified categories of technologies and techniques while demonstrating their practical applicability.

1.3 STRUCTURE OF THE DOCUMENT

Section 2 of the document discusses the relation between data protection engineering with data protection by design, data protection impact assessment and privacy enhancing technologies towards satisfying the overarching data protection principles. Section 3 discusses anonymisation and two prominent anonymisation techniques while also referring to past ENISA work in the area of pseudonymisation. Section 4 discusses some of the available techniques beyond encryption in the areas of data masking and privacy preserving computations while Section 5 discusses technologies on privacy preserving access control, storage and communications. Section 6 presents technical measures in the greater area of transparency, intervenability and user control tools while Section 7 concludes the document and provides recommendations for future work in the area.

[2] Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act) http://data.europa.eu/eli/reg/2019/881/oj

2. ENGINEERING DATA PROTECTION

2.1 FROM DATA PROTECTION BY DESIGN TO DATA PROTECTION ENGINEERING

The notion of privacy and data protection engineering has already been described in the past, either through a set of engineering strategies towards the principle of privacy by design or as a set of data protection goals.

Privacy and data protection engineering has already been described in the past, either through a set of engineering strategies towards the principle of privacy by design or as a set of data protection goals

For example, in its 2015 report [3], ENISA explored the concept of privacy by design following an engineering approach. Further to the analysis of the concept the report, using relevant work in the field, presented eight privacy by design strategies, both data oriented and process oriented, aimed at preserving certain privacy goals. As a different approach to privacy and data protection engineering, a framework comprising of six goals was proposed in order to identify safeguards for IT systems processing personal data. In addition to the typical security triad of "confidentiality", "integrity" and "availability", three additional goals "unlinkability", "transparency" and "intervenability" were also proposed. Significant work in the area of privacy engineering was also published in [4] & [5] and by the EU funded research project PRIPARE[3] . Despite the different starting points of each approach, they all put forward proposals for linking data protection requirements with technical requirements by either a methodology, specific goals or a set of strategies that had to be adhered to.

In its Preliminary Opinion 5/2018 [6] on privacy by design, the European Data Protection Supervisor (EDPS) provided a detailed overview of privacy engineering methodologies as a mean to translate the principles of privacy by design and by default. In the same Preliminary Opinion, the EDPS provided examples of methodologies to identify privacy and data protection requirements and integrate them into privacy engineering processes with a view to implementing appropriate technological and organisational safeguards. Some of these methodologies define data protection goals directly from privacy and data protection principles, such as those of the GDPR, or derive them from operational intermediate goals. Other methodologies are driven by risk management.

2.2 CONNECTION WITH DPIA

The Data Protection Impact Assessment (DPIA) is one of the requirements introduced under the GDPR and can be also perceived as part of the "protection by design and by default" approach. Further to the emphasis put by these principles on the engineering of data protection requirements into processing operations, such emphasis is also evident in Article 35 (7)(d) of the GDPR. The legislator explicitly mentions "*the measures envisaged to address the risks, including safeguards, security measures and mechanisms…*" which clearly extends beyond the traditional deployment of technical and organizational measures and calls for a more detailed analysis, selection and operation of techniques able to ensure the required level of protection. It is interesting to note that these provisions are also linked to the level of risk of personal data processing (which again works as a threshold for the adoption of relevant measures). This notion is also evident in some of the approaches proposed by Data Protection Authorities on

[3] http://pripareproject.eu/ . Relevant publication is available under https://www.slideshare.net/richard.claassens/pripare-methodologyhandbookfinalfeb242016

impact assessment such as Privacy Impact Assessment (PIA)[4] and relevant DPIA templates and guidance made available by other National Data Protection Authorities[5].

2.3 PRIVACY ENHANCING TECHNOLOGIES

Privacy Enhancing Technologies can be categorized based on the characteristics of the technology used in relation to the data being processed

Privacy Enhancing Technologies (PETs) cover the broader range of technologies that are designed to support implementation of data protection principles at a systemic and fundamental level. As described in [7], PETs are "*a coherent system of ICT measures that protects privacy by eliminating or reducing personal data or by preventing unnecessary and/or undesired processing of personal data, all without losing the functionality of the information system*". PETS, as technical solutions, can be perceived as building blocks towards meeting data protection principles and the obligations under GDPR Art. 25 on data protection by design. Therefore, they also comprise elements of the building blocks of data protection engineering.

As PETs can vary from a single technical tool to a whole deployment depending on the context, scope and the processing operation itself, it is evident that there is not a one-size fits all approach and there is a need for further categorization across different PETs. Towards this direction ENISA has put forward a methodology [8] on analysing the maturity of PETs and a framework on assessing and evaluating PETS in the context of online and mobile privacy tools. As highlighted by the Spanish Supervisory Authority (SA) in [9], a number of initiatives exists on classifications of PETs; either based on their technical characteristics or on the goals they pursue (in relation to the data protection principles they can support).

With regards to specific tools and technologies, another categorization can be based on the characteristics of the technology used in relation to the data being processed. More specifically, these characteristics can be:

- **Truth-preserving:** The objective of privacy engineering is to preserve the accuracy of data while reducing their identification power. This goal can be achieved for instance diluting the granularity of data (e.g. from date of birth to age). In this way data are still accurate but in a "minimized way", adequate for the purpose at stake. Also, encryption may be regarded as a truth preserving technique, since encryption applied in the reverse direction fully restores the original data without injecting any uncertainty in the process
- **Intelligibility-preserving:** Data are kept in a format which "has a meaning" for the controller, without disclosing real data subjects' attributes. For instance, the trick of introducing an offset to a hospitalization date keeps the day/month/year format but breaks the link with the true data of an identified patient. Also, the injection of noise is an intelligibility preserving techniques since it does not alter the look-and-feel of data providing confidentiality safeguard on the true data.
- **Operable Technology:** Mathematic and logic operations (e.g. a sum or a comparison) can be executed on the results of their applications. Operability does not necessarily entail intelligibility, since (as it will be said in this report) there are families of encryption techniques in which the (non-intelligible) results are directly operable using operations that are correctly executable in the encrypted domain.

Further to these characteristics, additional categorization can be performed with regards to the GDPR data protection principles that each category can support, at least in theory. Attempting to perform such a taxonomy could be of great value to data controllers and processors as it would provide a reference model of either what purposes each tool or technique can serve or as an indication of what is already achieved by already deployed tools and techniques. It should be noted however that the overall analysis should always be performed per processing operation and also include aspects such as nature, scope, context and purposes of processing, similar to the notion of DPIA.

[4] https://www.cnil.fr/en/privacy-impact-assessment-pia
[5] https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/vurdere-personvernkonsekvenser/vurdering-av-personvernkonsekvenser/

3. ANONYMISATION AND PSEUDONYMISATION

A generic anonymisation scheme cannot be applied to all use cases and provide full and unlimited protection

Anonymisation and pseudonymisation are two very well-known techniques that are widely used to practically implement data protection principles such as data minimization. Pseudonymisation is also explicitly mentioned in the GDPR as technique that can support data protection by design (Art. 25 GDPR) and security of personal data processing (Art. 32 GDPR). However, they are often confused, when in fact there is an important difference among them and their application in practice. As already pointed out by Working Party 29 [10] and according to GDPR Recital (26), anonymous information refers to information which does not relate to an identified or identifiable natural person - and, thus, anonymous data are not considered as personal data. On the contrary, according to Art. 4 (5) pseudonymised data, which can be (re)attributed to a natural person with the use of additional information, are personal data and GDPR data protection principles apply to them. A common mistake is to consider pseudonymised data to be equivalent to anonymized data.

In the area of pseudonymisation, ENISA has published over the last years a number of reports [11], [12] & [13] that cover the notion and role of the technique under the GDPR, different pseudonymisation techniques and models and a number of use cases where its applicability is demonstrated in practice. To this end, the focus of this Section is primarily on data anonymisation, aiming to discuss briefly k-anonymity and differential privacy as two possible techniques to anonymise relational or tabular data.

In the case of non-tabular data or sequential data, anonymisation might not be as easy or as straight-forward. For example in the case of mobility data, relevant studies [14], [15] & [16] have proved that by knowing 3 or 4 spatial-temporal points of a trajectory was sufficient to re-identify, with a high probability, a person in a population of several million individuals. Possible solutions can be to publish only statistics on different trajectories propose or publish synthetic data, i.e., artificially generated trajectories from the statistical characteristics of real trajectories [17]. Synthetic data are discussed further in Section 4.5.

3.1 ANONYMISATION

Data anonymisation is an optimization problem between two conflicting parameters: data utility and re-identification protection. In fact, data anonymisation is achieved by altering the data, either by noising or by generalizing. Providing strong re-identification protection usually requires to strongly alter the dataset and therefore negatively impact its utility. Data anonymisation therefore entails finding the best trade-off between these two parameters; and this trade-off often depends on the application and the context (i.e., how the dataset is distributed and used). As also mentioned in [10], [18] & [19], we should not assume that a generic anonymisation scheme can be applied to all use cases and such a scheme will be able to provide full and unlimited protection. Each solution must be adapted according to the type of data, the processing operation, the context and the possible attack models. This notion should be considered as applicable to every technique and technology discussed within this document. As mentioned in the Working Party 29 Opinion [10], "*An effective anonymisation solution prevents all parties from singling out an individual in a dataset, from linking two records within a dataset (or between two separate datasets) and from inferring any information in such dataset.*"

The next two sections provide a quick overview of the two most popular anonymisation approaches, namely k-anonymity and ε-differential privacy. A more in-depth overview of existing anonymisation schemes can be found in [10], [20] & [21].

3.2 k-ANONYMITY

The k-anonymity model was introduced in the early 2000s and it is built on the idea that by combining sets of data with similar attributes, identifying information about any one of the individuals contributing to that data can be obscured. As discussed in [22] a dataset is considered to provide k-anonymity protection if the information for each data subject contained in the dataset cannot be distinguished from at least k-1 data subjects whose information also appears in the dataset. The key concept is to address the risk of re-identification of anonymized data through linkage to other available datasets. For example, a sample data set is presented in Table 1 below.

k-anonymity is built on the idea that by combining sets of data with similar attributes, identifying information about any of the individuals can be obscured

Table 1: Initial Dataset

Name	Gender	Zip Code	Year of Birth	Diagnosed Medical Condition
George S.	M	75016	1968	Depression
Martin M.	M	75015	1970	Diabetes
Marie J.	F	69100	1945	Heart rhythm disorders
Claire M.	F	69100	1950	Multiple sclerosis
Amelia F.	F	75016	1968	Nothing
Annes J.	F	75012	1964	Rheumatoid arthritis
Sophia C.	F	75013	1964	Blood Disorder
Simon P.	M	75019	1977	Sarcoidosis
Michael J.	M	75018	1976	Lymphoma

To anonymise the data of Table A, several techniques are possible such as deletion or generalization[6]. In this example, the attribute *Gender* was kept unmodified as it was considered important for the study of medical conditions. Furthermore, the *Zip Code* of the user's address and the *Year of Birth* attributes were generalized by retaining only the department zip code and by using intervals of 10 years, respectively. Attempting to k-anonymize the data with a *k* value of two (2) and with respect to the quasi-identifier {Zip Code, Year of Birth, Gender} the initial data set is transformed to table, because for each triplet of values, there are at least two entries in the table corresponding to it, as presented in Table 2 below.

Table 2: k-anonymised data (with *k*=2)

Zip Code	Year of Birth	Gender	Diagnosed Medical Condition
75***	[1960-1970]	M	Depression
			Diabetes
69***	[1940-1950]	F	Heart rhythm disorders
			Multiple sclerosis
75***	[1960-1970]	F	Nothing
			Rheumatoid arthritis
			Blood Disorder
75***	[1970-1980]	M	Sarcoidosis
			Lymphoma

[6] Generalization can also be achieved by deletion of an attribute (column)

K-anonymity suffers from several limitations. For example, the k-anonymity criterion does not protect against homogeneity attacks, where all the records grouped in an equivalence class have the same or similarly sensitive value. Various extensions to the k-anonymity model have been introduced to address this issue, such as l-diversity, which ensures that for each quasi-identifier value corresponding to *k* data, there will be at least l representative values for the sensitive data [23] & [24]. Table 2 is 2-anonymous and 2-diverse, because there are always at least two different medical conditions within a group of individuals with the same quasi-identifier. However, if Simon P. had Lymphoma instead of Sarcoidosis, Table 2 would still be 2-anonymous, but would no longer be 2-diverse. In this case, one could infer that Michael J. who belongs to the group defined by (75, [1970-1980], M) has Lymphoma, whereas before this prediction was possible only with a probability of 50% (1/2). Another weakness of k-anonymity is that it does not compose, i.e., several k-anonymized datasets of the same individuals may be combined to re-identified individuals [25]. It is therefore very difficult to give an a priori guarantee on the risk of re-identification, which might depend of the adversary's knowledge.

The protection guarantee in k-anonymity depends on the value of k. intuitively, a large k value provides better protection than a smaller value, at the cost of data utility. To select a parameter for a privacy definition, the link between the parameter value and the risk of a privacy incident happening needs to be known. As shown previously, estimating quantitatively such risk, and therefore the corresponding k value, in k-anonymity is very difficult [26]. In the healthcare domain, when medical data is shared with a small number of people (typically for research purposes), *k* is sometimes chosen between 5 and 15 [27]. However, this choice is very arbitrary and ad hoc.

3.3 DIFFERENTIAL PRIVACY

Differential privacy (DP) algorithms [28] can provide assurance that after analyzing a dataset of several individuals, the outcome of the analysis will not be affected and will remain the same, even if any individual's data (up to ε) was not included in the dataset. In other words, differential privacy allows to study larger statistical trends in the dataset but protects data about individuals who participate in the dataset. Learning such trends (i.e., inferences which are generalizable to a larger population in interest) is probably the ultimate goal of any data release in general. Differential privacy is not an anonymisation technique per se, but a model on which anonymisation techniques can be developed and allows to quantify a risk of re-identification.

Differential privacy allows to study larger statistical trends in a dataset but protects data about individuals who participate in this dataset

To make a process differentially private, it has to be modified a little bit by, typically, some randomness or noise. This noise has to be calibrated to the value ε and the sensitivity, i.e., how much any individual contributes to the result of the process. Although there is still no rigorous method for choosing the key parameter ε [29], most systems implementing DP choose a value of ε close to 1. In any case, one must choose the smallest possible one that leads to acceptable utility and privacy. For example, let's assume that a city wants to publish the number of individuals that suffer from a specific chronic disease. This publication can take the form of a histogram where each bucket corresponds to the count of suffering individuals in a given district. Each individual can at most influence one of these buckets by the value 1 (an individual suffers or does not suffer from the disease and only live in one district). The sensitivity is then 1. It is known that this release can be made ε-differentially private by just adding Laplace noise of scale 1/ε to each bucket count [30]. The effectiveness of protection is determined in large part by that factor ε, so selecting its optimum value must be done in the context of data, the size of the population of users in the dataset and the processing that is being undertaken.

It is noteworthy that DP-based anonymisation can come in two flavours: global or local anonymisation. In the global model of differential privacy, data are collected by a central aggregator that transforms it, typically by adding noise, with a differentially private mechanism. This model requires to fully trust the aggregator. Instead in the local model, the participating users apply a differentially private mechanism to their own data before sending it to the aggregator. As a result, the aggregator does not need to be trusted anymore. Usually, the local

model requires to add more noise, and therefore reduces accuracy, although the use of secure aggregation techniques can sometime be used to minimize this accuracy degradation [31].

One of the main benefits of Differential Privacy is that the privacy loss can be quantified, even if a given dataset is anonymized several times for different purposes or different entities (we say that “Differential Privacy composes”). For example, the same dataset that is anonymized twice (for example by 2 different entities), each with a privacy value of ε, is still differentially private but with a privacy parameter of 2 ε. Another important propriety of differential privacy is that post-processing is allowed. In other words, the result of the processing of differential private data through a fixed transformation remains differential private.

3.4 SELECTING THE ANONYMISATION SCHEME

Data anonymisation is a complex process that should be performed on a case-by-case basis. Possible solutions depend on many parameters that vary from one application to another, such as the type of data (temporal, sequential, tabular data, etc), the data sensitivity or the acceptable risk and performance degradation levels. Any anonymisation procedure should be combined with a risk-benefit analysis that defines the acceptable risk and performance levels. This risk analysis will guide the data controller in selecting the model, the algorithm and the parameters to use.

Anonymisation procedures should be combined with a risk-benefit analysis that defines the acceptable risk and performance levels

The adopted anonymisation solution should also depend on the context, for example how the anonymised dataset will be distributed. A “release-and-forget” model, where anonymised data is publicly released without control, requires stronger protection than an “Enclave model”, where the anonymised data is kept by the data controller and only queries can be performed by qualified researchers. k-anonymity and Differentially Privacy approaches are sometimes perceived as competing approaches, and that one can be used instead of the other. However, these approaches are quite complementary and are adapted to different applications as discussed in [32]. k-anonymity is easy to understand and is well adapted to tabular data. It is better suited to publish anonymised data that can be used for different purposes. However, as it is vulnerable to a number of attacks and its security relies on the adversary background knowledge, it is not advisable to use it in a “release-and-forget” mode.

The Differentially Privacy model provides stronger protection than k-anonymity due to the added randomness that is independent of the adversarial knowledge. As opposed to k-anonymity, Differential Privacy does not need attack modelling and is secure no matter what the attacker knows. It is therefore better adapted to the “release-and-forget” mode of publication. However, DP is not well adapted for tabular data but more suited for releasing aggregated statistical information (counting queries, average values, etc.) about a dataset. Furthermore, DP-based anonymisation scheme often needs to be tailored to the data usage. It is challenging to generate a DP-anonymised dataset that provides strong protection and good utility for different purposes [33]. Furthermore, DP provides better performance for datasets where the number of participants is large but each individual contribution is rather limited.

4. DATA MASKING AND PRIVACY-PRESERVING COMPUTATIONS

Masking is a broad term which refers to functions that when applied to data, they hide their true value. The most prominent examples are encryption and hashing but as the term is rather broad, it also covers additional techniques, some of which will be discussed within this section. The main usability of masking with regards to data protection principles is integrity and confidentiality (security) and depending on the technique or the context of the processing operation it can also include accountability and purpose limitation.

4.1 HOMOMORPHIC ENCRYPTION

Homomorphic encryption is a building block for many privacy enhancing technologies like secure multi-party computation, private data aggregation, pseudonymisation or federated machine learning to name a few. Homomorphic encryption allows computations on encrypted data to be performed, without having to decrypt them first. The typical use case for homomorphic encryption is when a data subject wants to outsource the processing of her personal data without revealing the personal data in plaintext. It is apparent that such functionalities are very well suited when processing is performed by a third party such as a cloud service provider.

Homomorphic encryption allows computations on encrypted data to be performed, without having to decrypt them first

There are two types of homomorphic encryption: partially and fully [34]. Partially Homomorphic Encryption (PHE) is where only a single operation can be performed on cipher text, for example, addition or multiplication. Fully Homomorphic Encryption (FHE) on the other hand can support multipliable operations (currently addition and multiplication), allowing more computation to be performed over encrypted data. Homomorphic encryption is currently a balancing act between utility, protection, and performance. FHE has good protection and utility but poor performance. PHE on the other hand has good performance and protection, but its utility is very limited. There is however a catch; the performance of FHE is quite inefficient, where simple operations can take anywhere from seconds to hours depending on security parameters [35].

The choice of the homomorphic encryption depends on the desired level of protection in combination with the complexity of the computations to be performed over the encrypted data. If the operations are complex, the encryption scheme will be more expensive. The complexity of the computation is not measured as it is done classically in computer science (time and memory) but it is measured by the diversity of operations (addition and multiplication) performed on the inputs. If the computation only requires addition (like in the sum of some values) then partially homomorphic encryption can be used. If the computation requires some addition and a limited number of multiplications then somewhat homomorphic encryption, which is similar to partially homomorphic encryption but with a limitation on the number of operations instead of the types of operations, can be used. If the computation requires many additions and multiplications, then fully homomorphic encryption needs to be used.

4.2 SECURE MULTIPARTY COMPUTATION

The concept of secure multiparty computation (SMPC) refers to a family of cryptographic protocols that was introduced in 1986 and attempts to solve problems of mutual trust among a set of parties by distributing a computation across these parties where no individual party can see the other parties' data. Protocols of secure two-party computation, can for example

calculate functions over the input data of two parties, without revealing the input data of one party to the other party. Prominent variations of SMPC include the Byzantine Agreement [36] where the computation extends to multiple parties and auctioning [37] where participants can place bids for an auction without revealing their bids. The latter one is already deployed as a real-ife application[7] in Denmark where Danish farmers determine sugar beet prices among themselves without the need for a central auctioneer.

Secure multiparty computation attempts to solve problems of mutual trust among a set of parties where no individual party can see the other parties' data

The most prominent example of SMPC is found in blockchain technology: a set of parties, called "miners", have to determine and agree upon the next block to append to the blockchain ledger. This problem can be separated into two sub-tasks:

a) Determine a valid block (or set of blocks) to append to the blockchain, and
b) Agree with all other miners on the blockchain that this new block is the one to be appended.

Task a) can be done individually by each miner. This task consists in finding a valid hash value that fulfils a set of requirements by means of brute-force search (this is actually the power-consumptive part of the blockchain technology), and does not involve multiple parties yet. Once a miner finds and announces such a valid hash value, the majority of miners have to reach consensus that this hash – and its new block of transactions – is to be appended to the block chain. This task is similar to Lamport's byzantine agreement protocol, as some miners might play false, or may propose a different hash value and block that also satisfies all requirements.

In general, secure multi-party computation protocols exist for every function that can be computed among a set of parties. In other words, if there is a way that a set of parties can jointly compute the output of the function (by exchanging some messages and calculating some local intermediate results), then there always exists a secure multi-party protocol that solves that very problem with the security guarantees required. Unfortunately, in many cases, such a secure multi-party computation protocol can become very complex in application, and may easily demand a huge network communication overhead. Hence, it may not be suited for application scenarios with rapid real-time requirements.

Depending on the exact protocol chosen, secure multi-party protocols support the privacy protection goals of confidentiality (as the inputs of other parties are not revealed) and integrity (as even inside or external attackers cannot easily change the protocol output). This distributes the total power among all parties involved, which can be a huge number of entities in real-world applications like blockchain. Thereby, it becomes unrealistic that any individual party may decide and enforce its decision unilaterally against the other parties.

Moreover, given that the utilized secure multi-party protocol must be known to each party involved, this approach fosters transparency as to what type of processing is applied to the input data. On the downside of this approach, it is far from trivial to manually override the result of a secure multi-party computation in case of errors. If, for instance, an e-mail address is written to a block of the blockchain, and its hosting block is agreed upon by the consensus protocol among the miners, it becomes a part of the blockchain forever. Removal of this e-mail address from the blockchain later on is almost infeasible, as it would require every miner to locally remove it from the block, and ignore the error this deletion causes to the hash values in the modified blockchain – a direct violation of the blockchain protocol.

4.3 TRUSTED EXECUTION ENVIRONMENTS

Encryption is a powerful tool to protect data, however it becomes unusable if the device that is used to store, encrypt or decrypt the data is compromised. In this case the adversary can get access to the decryption materials and to the plaintext data. Trusted execution environment

[7] https://partisia.com/better-market-solutions/mpc-goes-live/

(TEE) can play a key role in protecting personal data by preventing unauthorized access, data breaches and use of malware. It provides protection against strong adversaries that get access, either physically or remotely, to the devices. With a TEE, the processing of the data takes place internally in the enclave. It is then theoretically impossible to obtain any data.

A trusted execution environment (TEE) is a tamper-resistant processing environment on the main processor of a device. Running parallel to the operating system and using both hardware and software, a TEE is intended to be more secure than the traditional processing environment. This is sometimes referred to as a rich operating system execution environment, or REE, where the device OS and applications run. It guarantees the authenticity of the executed code, the integrity of the runtime states (e.g. CPU registers, memory and Input/Output), and the confidentiality of its code, data and runtime states. The TEE resists against software attacks as well as the physical attacks performed on the main memory of the system. Unlike dedicated hardware coprocessors, TEE is able to easily manage its content by installing or updating its code and data.

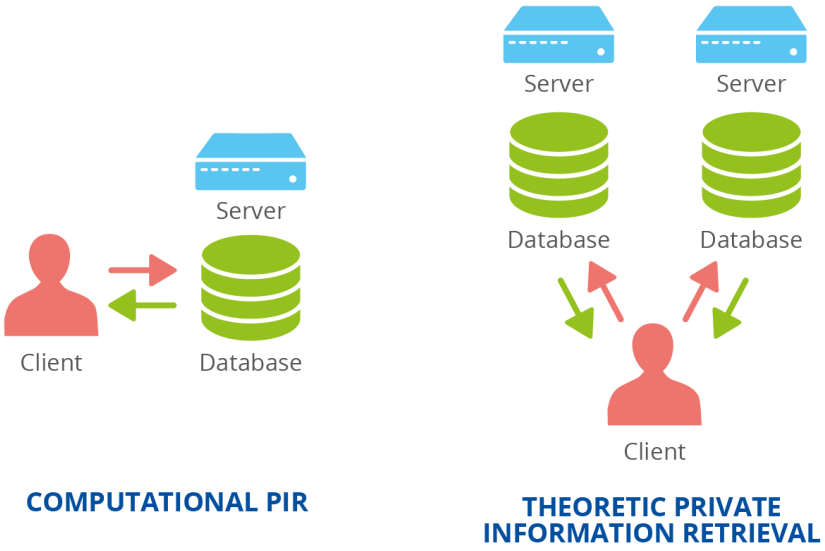
TEEs are used widely in various devices, such as smartphones, tablets and IoT devices. TEEs can also play an important role to secure servers. They can execute key functions such as secure aggregation or encryption to limit the server's access to raw data. It may provide opportunities to provide verifiable computations and increase trust. Indeed, TEEs enable clients to attest and verify the code running on a given server. In particular, when the verifier knows which binary code should run in the secure enclaves, TEEs can be used to verify that a device is running the correct code (code integrity). For example, in a federated learning setting, TEEs and remote attestations may be particularly helpful for clients to be able to efficiently verify key functions running on the server, such as secure aggregation or shuffling.

4.4 PRIVATE INFORMATION RETRIEVAL

Private information retrieval allows a user to recover an entry in a database without which element was queried

Private information retrieval (PIR) is a cryptographic technique which allows a user to recover an entry in a database without revealing to the data custodian (e.g. the database owner or administrator) which element has been queried [38]. This is why it can be used as a data minimization technique by data controllers. Let's assume that a company wants to provide access to a database to its customers. In a default setting, each time a customer makes an access to the database, the data custodian knows which entry has been accessed. Over time, the data controller will be able to identify which database entries are of interest to the customers. By implementing private information retrieval, the data controller minimizes the amount of information revealed on what was accessed PIR prevents the data controller from learning which entries have been accessed.

Figure 1: Private Information Retrieval Models



There are two main models of private information retrieval. The first model is Computational Private Information Retrieval and there is only one server storing the database. This model is considered to provide better level of protection but has limitations with regards to the connections that can be established to the server and the database. In the second model, Information Theoretic Private Information Retrieval, the database is stored on several servers which are controlled by different owners. This model allows for better communication complexity but it is assumed that the servers do not collude or exchange information. Additional information on PIR are available in [39] and [40].

4.5 SYNTHETIC DATA

Synthetic data is a new area of data processing in which data are elaborated in a way that they realistically resemble real data (both personal and non-personal), but actually they do not refer to any specific identified or identifiable individual, or to the real measure of an observable parameter in the case of non-personal data. For example, they may be entirely simulated data with the goal of testing services or software applications. Alternatively, they can be personal data, which are manipulated in a way to limit the potentials for individuals' re-identification. By the term synthetic data, it is also possible to refer to the combination of multiple data sources in order to have better estimates of a population's parameters (a sort of cross-enrichment between different datasets).

Synthetic data is data elaborated in a way that they realistically resemble real data, but actually do not refer to any specific identified or identifiable individual

By using synthetic data, a controller will be respectful of individuals' confidentiality, since they differ from real data and the generation and processing of synthetic data does not invade the personal sphere of data subjects (in particular when real data refer to individuals' sensitive characteristics, or to rare attributes that may be difficult to retrieve or may have a significant power of identification). But at the same time this choice may introduce issues in terms of data accuracy. So, controllers will always need to reconcile a tension between different data protection principles, especially if the result of the processing entails consequences (i.e. legal or health consequences) for data subjects. Synthetic data should always be adopted having in mind the necessity to explore with a trial-and-error approach whether their use actually generates more accurate and unbiased estimates over time. In this perspective, synthetic data are privacy engineering tools that can provide granular data without sacrificing data subjects' privacy and confidentiality.

Many practical alternatives exist for generating synthetic data. The easiest option is drawing samples from a known distribution. In this case, the outcome does not contain any original (and

personal) data and re-identification is an unlikely occurrence, mainly due to randomness. More complex options rely on mixing real data and fake ones (the latter being still sampled from known multivariate distributions, conditioned on the real observed data). In this case, some disclosure of personal data and re-identification is possible due to the presence of true values within the dataset. The practical generation of synthetic data today, due to the variety of attributes involved and the oddness of the probability distributions, is based both on the use of classical random number generation routines but also, more and more, on the application of artificial intelligence and machine learning tools.

There are pros and cons in the use of synthetic data and controllers must be aware of both. On the benefits side, synthetic data are machine generated data, and as such they are easy and almost costless to reproduce. The burden of collection is voided for controllers, as well as the intrusiveness for data subjects. Furthermore, synthetic data can also cover situations in which it may very difficult or even unethical to collect (personal) data. For instance, in counterfactual analyses where the goal is to study the causal effects of a specific intervention and implementing this intervention may not be a practical option. Think of situations in which one is interested in the effect of a new treatment on a pathology, or the consequences of an exposure to a risk factor for human health. In all those circumstances it may not be possible to give the new treatment to the entire population,

5. ACCESS. COMMUNICATION AND STORAGE

5.1 COMMUNICATION CHANNELS

Secure communication channels, as the name implies, relate to providing secure exchange of data among two or more communicating parties. They are usually designed to enhance communications' privacy in a way that no unauthorized third party can access the content and in some cases the participants or even the metadata of the communication taking place. The GDPR requires the security of the processed personal data, including during transmission, pursuant to recital 49 and article 32 of the GDPR. Protection of privacy in the electronic communications sector and processing of metadata, such as traffic data, is also covered by the ePrivacy Directive[8].

Communication channels should go beyond the provision of security as their core functionality and incorporate additional privacy enhancing characteristics

From the data protection engineering perspective, communication channels should go beyond the provision of security as their core functionality and incorporate additional privacy enhancing characteristics, such as who can have access to the content of the communication, including the providers, location and access to the encryption keys, location and type of the provider, user information disclosed etc. Towards this direction, two technologies are being discussed below, namely End-to-End encryption and proxy routing.

5.1.1 End to End Encryption

End-to-End Encryption (E2EE) is a method of encrypting data and keeping them encrypted at all times between two or more communicating parties. Only the parties involved in the communication have access to the decryption keys. The implementation of end-to-end encryption is clearly a fundamental feature for secure messaging apps and has gained a lot of traction during recent years, where a number of widely used Over-The-Top (OTT) services such as messaging apps claim to implement End-to-End encryption. The main difference between E2EE and link encryption or encryption in transit is that in the latter cases, the content can be accessed by the server, depending on where the encryption takes place or where the encryption keys are stored. In a typical E2EE scenario, the encryption keys are stored at the end user devices and the server has information only on communication metadata (communication participants, date/time, etc). However, E2EE is reliable only until one the endpoints is compromised. An overview of end to end encrypted message protocols is available in [41] & [42].

Following the judgment C-311/18 (Schrems II)[9], which concerned the personal data transfer of an EU citizen to the US, the EDPB published its recommendation [43] on measures that supplement transfer tools to ensure compliance with the EU level of protection of personal data. Under Use case 3 and the conditions mentioned thereto, end to end encryption, combined with transport layer encryption, is considered as a mean to allow personal data transfers to non-EU countries for specific scenarios.

[8] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) , revised by Directive 2009/136/EC https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02009L0136-20201221

[9] Judgment of 16 July 2020, *Schrems,* C-311/18, EU:C:2020:559, https://curia.europa.eu/juris/document/document.jsf?text=&docid=228677&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=40128973

5.1.2 Proxy & Onion Routing

Further to the content of communications discussed earlier, there is also the aspect of communication metadata (data that describes other data and include information on who, what, where, when, etc,) which according to the EDPB statement [44] on the revision of the of the ePrivacy Directive and the Proposal on ePrivacy Regulation[10] "*can allow very precise conclusions concerning the private lives of the people to be drawn, implying high risks for their rights and freedoms*". A more formal definition of electronic communications metadata is provided in Article 4(3)(c) of the proposal.

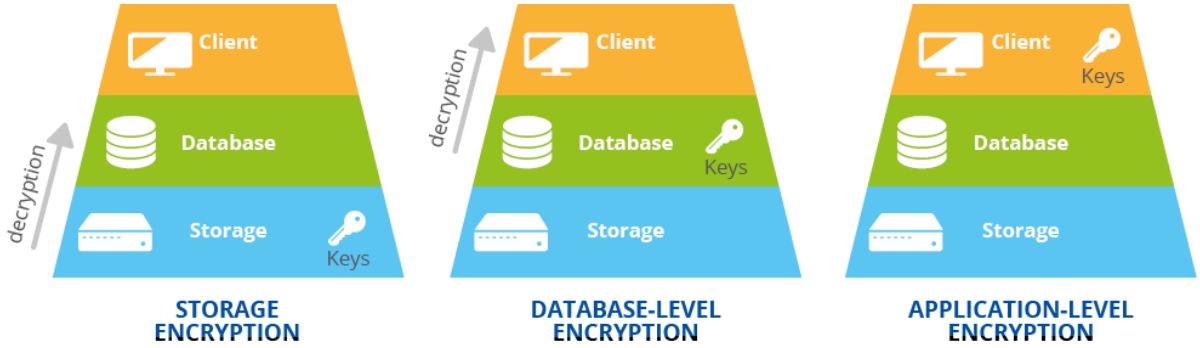
One possible model to protect metadata is the use of an onion routing network (e.g. Tor[11]) which supports anonymous communication over public networks. In onion routing user traffic is routed through a series of relay servers. [45], and each relay server receives layered encrypted data without knowing neither the original sender nor the final recipient. Such information is available only to the entry and exit node [46]. However, Tor is vulnerable to attackers who can observe traffic going in the entry and out of exit nodes and correlate messages, as discussed in [47].

5.2 PRIVACY PRESERVING STORAGE

Privacy preserving storage protects the confidentiality of personal data at rest and informs data controllers in case of a breach

Privacy preserving storage has two goals: protecting the confidentiality of personal data at rest and informing data controllers in case a breach occurs. Encryption is the main technique used to protect the data confidentiality from unauthorized access. Depending on the constraints of data controllers, it can be applied at three different levels: (i) storage-level, (ii) database-level and (iii) application-level encryption.

Figure 2: Database Encryption Options



File system and disk level encryption mitigate the risks of an intruder getting physical access to the disk storing the database. This approach has the advantage to be transparent for the users of the database however it is an all-or-nothing approach because it is not possible to encrypt only certain parts of the database and it is not possible to have a better granularity than at file level. In this solution, there is only one encryption key that is managed by the system administrators of the database. This key is held on the server hosting the database and must be protected by the highest privileged access.

Encryption can also be done at the database level. This approach provides more flexibility than the previous solution and can be applied at different granularities tables, entries or fields. It can also be applied when some data fields/attributes are more sensitive than others (political or religious belief for instance). However, because the encryption keys need to be stored with the

[10] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017PC0010
[11] Tor Project https://www.torproject.org/

database, an adversary who can connect to the server hosting the database can used forensics tool like Volatility[12] to recover the keys directly from the volatile memory.

In application-level encryption, all data are encrypted by the client with its own encryption keys and then stored. However, if several entries of the database are to be shared by different clients, the cryptographic keys need to be exchanged, which can jeopardize their security. It is possible to avoid this issue if specific encryption schemes are used, e.g. homomorphic encryption. The encryption keys do not need to be shared anymore as it is possible to perform computations over the encrypted data.

With regards to notification, canaries are a well-known security mechanism which is used to detect software attacks and buffer overflows. The concept of a canary can be transposed to personal data protection. Injecting a canary into a database entails inserting fake values in it which are not supposed to be used by anyone. Access of these values must be then monitored in order to detect a data breach. It is also important to notice that these fake values must not be distinguishable from the real ones. A possible implementation of such scheme will have a server storing the database and a distinct server which handles the requests to the database. The server handling the requests to the database must have the capability to identify requests for canaries, thus detecting a possible attack or breach. This model is particularly suitable for data controllers who want to use third party cloud-based storage. However, the data controller needs to find a good balance between the number of real entries in the database and the number of canaries (fake entries) and in any case, such techniques cannot be considered as panacea for timely identification of data breaches.

5.3 PRIVACY-ENHANCING ACCESS CONTROL, AUTHORIZATION AND AUTHENTICATION

Authentication, authorization and access control aim to prevent unauthorized and/or unwanted activity from occurring by implementing controls and restrictions on what users can do, which resources they can access and what functions they are allowed to perform on the data, including unauthorized viewing, modification, or copying. Authentication confirms the identity of a user requesting to access the data while authorization determines which actions an authenticated user can take. Access control refers to a technique which ensures that only authenticated users can access the information they are entitled to. These three elements are closely related and omitting even one of them can weaken the level of protection to the data, as authorized users can gain access to them or authorized users can perform unauthorized actions.

Zero-Knowledge proofs allow a user to prove that she knows a secret information without revealing anything on this secret

Depending on the context and the needs, certain access control mechanism seems to be more suitable than others. As also discussed in [48] , in a scenario where processing of customers' personal data for marketing purposes takes place through an on-line cloud storage provider, Discretionary Access Control (DAC) can be used for accessing data for a specific service request such as a print and delivery service. Through DAC, an employee is able to specify which data, for each, external to the organization, user, and what action(s) are permitted. DAC provides users with advanced flexibility on setting up desired access control properties, however on the negative side it relies heavily on user's awareness and understanding of associated risk. On the other hand, in a hospital information system, where each actor (doctor, nurse, administrative personnel) is assigned to different roles with different privileges (e.g. a doctor can access penitents' medical data), Role Based Access Control (RBAC) seems to be more appropriate.

[12] https://www.volatilityfoundation.org

5.3.1 Privacy-enhancing attribute-based credentials

Attribute Based Credentials (ABC) allow the authentication of an entity by selectively authenticating different attributes without revealing additional information that are typically used and could very well include personal data. For example, in order for a service provider to allow access to an online service, the provider needs to verify the age of the individual requesting access. Instead of asking for individual's age, the provider could ask for the value of an attribute which indicates whether the data subject is over 18 or not. Towards including even more privacy characteristics, Privacy-enhancing Attribute-Based Credentials (PABCs) [49] have been proposed. This technique authenticates the user through attributes in a data-minimizing way, by providing un-linkable (to each other) attributes. Notable research work in the area was performed by the ABC4Trust [50] H2020 research project and since then there is a number of deployments such as the IRMA card[13]. Recently, Decode[14] H2020 research project has implemented pilot demonstrations of attribute-based credentials exploitation in real life scenarios.

5.3.2 Zero Knowledge Proof

Zero-Knowledge proofs [51] are cryptographic primitives which can be used to enforce the confidentiality and the data minimization principles of the GDPR. The core idea of a zero-knowledge proof is to allow a user (a data subject) prove to a server (data controller) that she knows a secret information without revealing anything on this secret ([52]). Zero-knowledge proofs are mostly used to implement authentication schemes and several protocols are proposed in ISO/IEC 9798-5[15].

Zero-Knowledge proofs not only enforce confidentiality but also, compared to other authentication schemes such as user name/password, they enforce the data minimization principle. In password-based authentication scheme, a user sets a password and shares this password with a server.

When the user wants to authenticate herself to the server, she needs to provide her password, which is then compared to the one stored on the server. An adversary who wants to impersonate a user can steal the password either from the user or from the server. In zero-knowledge proofs authentication scheme, this risk is limited only to the user because the server does not know the secret used by the user to authenticate. The technique minimizes the amount of information that the server knows about the user and consequently reduces the attack surface. Zero-knowledge proofs are also a building block for many secure multi party computation protocols.

There are two variants of zero-knowledge proofs: interactive [51] and non-interactive [53]. Interactive zero-knowledge proofs require several communications between the user and the server. Non-interactive zero-knowledge proofs do not require any communication at all. Non-interactive zero-knowledge proofs are very popular in blockchain applications.

[13] IRMA app https://irma.app/
[14] https://decodeproject.eu/
[15] ISO/IEC 9798-5:2009 Information technology — Security techniques — Entity authentication — Part 5: Mechanisms using zero-knowledge techniques https://www.iso.org/standard/50456.html

6. TRANSPARENCY, INTERVENABILITY AND USER CONTROL TOOLS

A key element in any data protection concept is the enablement of human individuals to exercise their data protection rights themselves. This involves both access to information on data processing (transparency) and the ability to influence processing of their personal information within the realm of a data controller or data processor (intervenability). In this respect, a multitude of approaches and topics emerged from the privacy research community that can help implementing these rights and correlated services at data processing institutions. In this chapter, we present a selection of the most relevant ones.

Providing accurate data protection information is not an easy task as simplification of information might be necessary but might also create misunderstanding

Transparency on data processing is not only demanded by the GDPR, but also necessary for individuals to understand why their personal data is collected and how it is processed, e.g. whether it is transferred to other parties. While system designers or data protection officers may be able to understand and even demand detailed information about the data processing systems and processes, most users are not able to grasp what is laid down in a technical specification or legal documents and may even be overwhelmed when being presented with basic information on personal data processing. (Article 13 & Article 14 GDPR).

6.1 PRIVACY POLICIES

In the online world, a well-known instrument for providing information to the users is the privacy policy (sometimes also called: data protection statement, data policy, privacy notice or alike).

The first recommendation from the Article 29 Data Protection Working Party on how to inform online users about data protection issues stems from their recommendations published in 2004 [54] and stresses the possibility of a multi-layered approach, starting with essential information and providing in additional layers – if desired from the user – further information. The layered approach is particularly helpful for presentation on mobile devices where it would be cumbersome, if not impossible at all to read a lengthy text with full information on processing of personal data.

In 2017, the Article 29 Data Protection Working Party published guidelines on transparency [55] which referred to the obligations laid down in the GDPR. In particular, the document explains the meaning of the requirement of Article 12 (1) s. 1 GDPR: "*The controller shall take appropriate measures to provide any information referred to in Articles 13 and 14 and any communication under Articles 15 to 22 and 34 relating to processing to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child*."

Providing accurate data protection information is not an easy task, because on the one hand simplification may be necessary so that an average person can understand the information, but on the other hand the simplification must not provoke misunderstanding. For natural language information, several metrics have been proposed to measure the complexity and the

comprehensibility, e.g. of newspapers or terms and conditions of insurances[16]. Such metrics can be employed by controllers for scrutinizing the understandability of their privacy policy even though data protection authorities have not yet recommended the use of a certain metrics. For specifically addressing children, the Information Commissioner's Office has published a code of practice for online services [56].

When designing the presentation of the privacy policy, the users' potential devices have to be considered, e.g. the size of the screen. Also, the information should be accessible and designed in a way that is compliant with assistive technologies so that people with disabilities are not excluded. In specific situations, textual information is not appropriate, e.g. in phone calls or in some contexts in a smart home or a connected car. Also, the entire HCI communication should be checked concerning the information on data processing given to the users. Checks and tests for comprehensibility (and the absence of dark patterns) could involve the data protection officers and people with usability knowledge.

It has to be noted that the provision of information is not limited to one basic document such as the privacy policy, but ad-hoc personalised information can also be given during the actual usage as it is designed in the human-computer interface (HCI). This kind of information can influence users in making up their minds concerning data-protection relevant matters (e.g. which data to post in social networks) or on giving or withdrawing consent.[17] In the report titled "Deceived by Design", the Norwegian Consumer Protection Council has pointed out that the HCI design of many applications and services is not neutral, but employs so-called "dark patterns" that nudge users towards disclosing more data and making precipitate decisions on data processing methods [57].

6.2 PRIVACY ICONS

Privacy icons can become a very good approach to support, or even substitute, textual information on personal data processing

Better comprehensibility may also be achieved if the information is conveyed not only by text that requires reading skills and effort, but also via graphical symbols (icons). Icons are a well-known method to support, or sometimes substitute, textual information. The GDPR has taken this possibility up in Art. 12 (7). Until today, there is no standardized icon set to be combined with Art. 13, 14 information, but there have been various proposals from the research community [58], [59], [60] & [61].

Art. 12 (7) also introduces also introduces a requirement for a machine-readability of the information conveyed by icons. The machine-readability could facilitate a better comprehension of the general meaning of the icon and getting further information on the data processing that is addressed by the icon. The advantages of machine-readability of a privacy policy as such (and not only the icons) encompasses the option of translating into a language preferred (and understood) by the user and automatic interpretation on the user's machine, potentially matching the information from the privacy policy with preferences or demands configured on the user side (cf. Section 6.4 on privacy preference signals).

Some controllers have already introduced self-designed icons that are presented in combination with their privacy policy. In the absence of a standardized icon set they may be appreciated by users. However, as soon as standardized solutions for icons and machine-readability as published, they should be used. As already stated, there are several open issues concerning privacy policies such as a lack of definition of good practice and the lack of standards concerning machine-readability or a defined icon set. Furthermore, new technology scenarios

[16] E.g. LIX formula on readability, developed by Carl Hugo Björnsson in 1971: LIX(text) = TotalWords/Sentences + (LongWords x 100)/TotalWords;
CFP (Content Function Ratio) formula on informativity: CFR(text) = AmountOfContentWordTags / AmountOfFunctionWordTags;
HIX (Hohenheimer Verständlichkeitsindex) formula on comprehensibility, developed by University of Hohenheim, based on Amstad formula, 1. Neue Wiener Sachtext-Formel, SMOG-Index and LIX, https://klartext.uni-hohenheim.de/hix

[17] Since consent means "freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;" (Art. 4 No. 11 GDPR), it requires sufficient information.

have not been well reflected, yet, e.g. sensor technologies with restricted or missing user interfaces or complex or dynamic and thereby hard to understand data processing systems that may encompass several controllers or steadily changing IT systems.

6.3 STICKY POLICIES

The purpose of (machine-readable) privacy policies can be extended to govern the data processing itself. There have been several proposals on policy information that is bound to the data items, e.g. by complementing meta information with the intent of describing important properties (source, recipients, purposes, date of processing etc.) or technically controls the allowed and not allowed data processing operations (access, transfer, erasure etc.). An advantage of sticky policies is the combination of the technical data processing organisation and the provision of transparency. If a data protection management system governs all data processing at the controller, changes in processing automatically invoke changes in the information in the policy. Similarly, restrictions laid down in the policy can automatically be guaranteed. Access rights, time restrictions or event-based triggers can be defined in a policy language that is both executable in the data processing system and translatable into natural language in the privacy policy.

Through privacy preference signals, users can express their privacy preferences in a machine-readable way

The most prominent proposal in this respect is the work on “sticky policies” [62] where policies are “stuck” to the data and also travel with them in case of data transfer. Cryptographic methods are being used to prevent recipients from ignoring the attached policies. However, all kinds of policies do not fully exclude the possibility of misuse of personal data. Currently, there are no standardized solutions of such machine-readable policies that also control data processing operations.

6.4 PRIVACY PREFERENCE SIGNALS

While data controllers provide privacy policies to inform about data processing and data protection aspects, this is not necessarily a one-way communication. Since the 1990s, possibilities for users as data subjects to express their privacy preferences in a machine-readable way have been discussed as presented in [63]. While each user in the online or offline world may express demands or wishes on how their personal data should be processed, data controllers usually cannot (and won’t) fulfil arbitrary demands. Instead, they will stick to pre-defined standardised processing operations. Communication of standardised machine-readable privacy preference signals from the user side can be interpreted by servers that support these technical standards. A first example of such approach, which became an obsolete standard in 2018 is the “Platform for Privacy Preferences Platform” specification [64] that would allow to express privacy policies on the server side and a complementing user-side language specification [65].

Following the idea of expressing privacy-relevant wishes or demands by the user in machine readable format, several languages and protocols were developed – mainly in research projects and prototypes.

While these languages and protocols tend to be comprehensive and often complex approaches with many features, for practical applications a more simplistic approach seemed to be expedient. A prominent example was the “Do not track”-Standard (DNT)[18] where users could express via an HTTP header field if they didn’t want to be tracked. “DNT = 1” means “This user prefers not to be tracked on this request.”, while “DNT = 0” stands for “This user prefers to allow tracking on this request.” A third possibility would be to refrain from sending a DNT header[19] because the user has not enabled this function. A deficiency of the DNT standard was the lack of a supporting legislation: If the question of “tracking” or “non-tracking” can only be expressed

[18] https://www.w3.org/2011/tracking-protection/
[19] https://www.w3.org/TR/tracking-dnt/

has a “preference” instead of a clear demand and if only “polite servers” react accordingly, this won’t help to achieve reliability and clarity for users or service providers.

A follow-up standard is called “Global Privacy Control” (GPC)[20]. It enables users to send a “do-not-sell-or-share” signal via their browser to a website in which the user is requesting that their data not be sold to or shared with any party other than the one the user intends to interact with, except as permitted by law. Since mid-2021 the GPC signal is regulated in the adapted California Consumer Privacy Act (CCPA)[21]. Users who want to express a “do-not-sell-or-share” signal can use one of the supported browsers or extensions. Under the European data protection regime, service providers are currently not forced to implement specific protocols that interpret users’ privacy preference signals,

In the era of Internet of Things (IOT), the role of machine-readable policies as well as privacy preference signals will become more significant. Providers of websites or web services should support standardised privacy preference signals and take into account and respect demands expressed by the users when deciding on processing of personal data. However, it has to be noted that Article 25 (2) of the GDPR requires data protection by default without the necessity of users to explicitly state if they don’t agree with processing of their personal data such as profiling, sharing or selling: The controller has to ensure “that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility.”

If users preferences, expressed in such a standardised technical way, cannot be fulfilled, for transparency reasons, controllers should inform them (e.g. in their privacy policy) why this is the case. For instance, in some countries there may be laws that require longer retention periods that the users would expect. In addition to standardised privacy preference signals, users’ browsers may contain privacy-tools as add-ons or have a specific configuration concerning tracking, data minimization of identifiers, or script blockers. In case a restrictive configuration may prevent correct functioning of a website or web service can be employed, the providers should inform users about potential limitations and offer at least basic functionalities to those privacy-aware users. Effectively this means to respect privacy demands by users, no matter whether they use one of the upcoming standards of privacy preference signals or other tools.

6.5 PRIVACY DASHBOARDS

Privacy dashboards can be used as a mechanism to enhance transparency and possibly also intervenability for data subjects. The objective of privacy dashboards is to give data subjects an overview on how their personal data is processed by a data controller. In contrast to the information provided in privacy policies that often provide rather abstract descriptions of processing operations, privacy dashboards can be used to show the actual personal data items that are accessible by the controller. This facilitates a better understanding for data subject which of their personal data are being processed. Often, it is also shown when and to whom which personal data are being disclosed, e.g. if personal data are being transferred to other organisations or if (and possibly for which purpose) a person has accessed specific personal data items. Data subjects may also check whether their personal data, as shown via the privacy dashboard, is outdated, wrong, incomplete or excessive or whether the disclosure to others is plausible or rather unexpected.

Privacy dashboards can provide users with an overview on how their personal data is being processed by a data controller

Today’s existing privacy dashboards are usually provided by controllers who also decide to what extent information is being presented there, how much explanation is given e.g. on potential risks and what options can be employed by the user to adapt settings or change or delete personal data. Users cannot expect to be informed on all kinds of data disclosures, e.g. if

[20] https://globalprivacycontrol.org/
[21] https://oag.ca.gov/privacy/ccpa#collapse7b

law enforcement or public authorities demand (in their jurisdiction lawful) access to the user's personal data, but prohibit the notification of the user. Also, information on data breaches may be excluded from the presentation. In case the data processing of the controller encompasses profiling, the privacy dashboard should clarify which personal data are being used for the user's profile and at best which information is being derived from the aggregated personal data.

Controllers should consider the implementation of usable privacy dashboards for data subjects that fulfil the requirements of the GDPR. As discussed in Section 5.3, a reliable authentication of the user is necessary to prevent the disclosure of personal data to unauthorised persons. If user-side privacy dashboards become distributed, controllers should check whether these tools could be supported as transparency-enhancing technologies.

In the following sections two types of privacy dashboards are described: services-side privacy dashboards and user-side privacy dashboards.

6.5.1 Services-side privacy dashboards

One of the first privacy dashboards offered by a company was the Google dashboard[22] that acts as a central access point for account holders to manage their privacy settings. For companies that employ targeted advertisement, such privacy dashboard may also play a role for explaining the choice of advertisement ("Why do I see this advertisement?") and on this basis for asking the user to potentially adjust the configured or derived categories of presumably interesting advertisements. This user-centric approach is to be criticised since it provides only partial transparency. For Example, it is usually not explained in detail how the assumptions on the user's interests were generated – and since the business model of personalised advertising does not necessarily fulfil the principle of data minimisation, but instead nudges users to provide more information that can be used for their profiling.

Privacy dashboards are also known as a functionality of citizen portals for services of the public sector. Probably the first country to provide a tool that presents an overview of personal data and who has accessed them is Estonia: The RIHA system[23] shows for each public data base and governmental information system which personal data are being stored for which purpose and who can access it. The Estonian citizens can see which officials have viewed their personal data. This information is gathered from access log files. Citizens can monitor the access activities which must not happen without a justified reason, as regulated by national law [66].

A similar functionality of a privacy dashboard called "Datenschutzcockpit" is planned for the German public sector as regulated in the "Onlinezugangsgesetz". A privacy dashboard function within a citizen portal does not necessarily mean that all personal information has to be permanently stored in a central database; it may be alternatively implemented as a central view on decentralised information.

6.5.2 User-side privacy dashboards

User-side privacy dashboards are applications that run under the control of the user, e.g. as a tool on the user's device. Such tools, called "Transparency-Enhancing Technologies" have been developed under research projects and aim to enhance user's transparency on their disclosure of personal data to different controllers, potentially under various pseudonyms and to support the user in the management of identities. An overview is given in [67].

This kind of privacy dashboards would profit from standardised machine-readable policies and privacy preference signals so the shown overview of personal data processing could be based on reliable information as provided by a controller. It is to be expected that the standardisation

[22] Google account: myaccount.google.com
[23] Riigi infosüsteemi haldussüsteem: https://www.riha.ee/

of machine-readable policies and their distribution in practice would lead to development of user-side privacy dashboards.

Privacy dashboards can be the means not only for presenting an overview of relevant information on processing of personal data but they can also offer functionality for changing privacy settings or for exercising data subjects' rights.

6.6 CONSENT MANAGEMENT

Consent management relates to managing consents from users for processing their personal data

Most of the existing web services on the Internet are operated by companies that base the processing on the legal grounds of consent. When registering to use such services, the user has to agree to the terms of use, thereby expressing knowledge of and consent to the data processing under the conditions outlined in these documents.

Unfortunately, web services tend to change frequently; new functions are added and old ones are phased out. New business partners are involved which may act as data processors. Each time this happens, it becomes necessary to validate whether the current terms of use documents cover the changes made to the processing operations.

In such cases, it becomes inevitable for the data controller to change its terms of use accordingly (and potentially its privacy policy and other documents alike), e.g. by including the new business partners in the list of data processors, or by adding the new purpose of data processing. Once this is done, new customers will read and agree upon the new terms of use, thereby expressing consent to these data processing rules. However, existing customers that already agreed to the old terms of use did not express consent to the changed processing rules and cannot easily withdraw their previous consent. This situation may easily become problematic.

If different customers have agreed to different versions of terms of use at different times, their legal basis for data processing may differ. Consent cannot be considered to be automatically granted for every change made to data processing. It is therefore necessary to explicitly ask existing customers to review the updated terms of use and provide their consent once again. Depending on the type and implementation of web service in consideration, this may become an issue of its own.

Either way, in a realistic scenario, it is inevitable to allow for different customers using the same service under different terms of use – and thereby consent coverage. Hence, it becomes inevitable as well to keep track of these differences, i.e., to record which customer operates under which data processing consent. This is commonly referred to as a central aspect of consent management.

6.7 CONSENT GATHERING

For browser-based web services, the de-facto approach for collecting consent is to display the text of the terms of use to the customer for reading, then adding a button at the bottom that declares "I have read and understood these terms of use". Once the user clicks the button, the consent is considered to explicitly have been given, and that information is recorded in the customer's user profile (or, in worst case, is implicitly recorded by the fact that a user profile is created and the user is allowed to login).

Unfortunately, this approach has several drawbacks:

- Users tend to click the button without reading and understanding the document ("consent fatigue" [68] & [69])
- Users with disabilities cannot read or understand the document
- Browser display issues may hinder users from reading the document

- Services that cannot be operated via browser cannot utilize this approach
- Services on embedded computers (e.g. cars, IoT devices) may not have a screen to show the terms of use document
- Services on embedded computers may not have a button nor other input device to express agreement

Nevertheless, consent gathering is required even in such circumstances, and the expression of consent must be given and documented validly, in order to be utilized as a legal basis for data processing.

6.8 CONSENT MANAGEMENT SYSTEMS

From an implementation perspective, there are multiple approaches for managing consent in dynamic real-world application scenarios (see e.g. [70], [71] & [69]). In the application domain of healthcare, where patients are asked for consent prior to medical operations, highly elaborated concepts for consent management have been proposed. For obvious reasons, doctors in such institutions have a pressing need to document such expression of consent prior to any operation, as lack of consent might render e.g. a surgical operation as inflicting body harm.

Here, multiple approaches for consent management have evolved, for documenting a one-time consent to a medical operation or treatment, but they are mostly based on a large legal text with a hand-written signature of the patient below. Digital equivalents utilize electronic documents, authentication tokens such as personal ID cards, and technologies like blockchain for permanently storing exact versions of consent documents along with the expression of consent of the users [72]. Here, we have a strong set of security requirements for gathering and documenting these electronic expressions of consent, e.g. with respect to integrity and availability of the consent forms.

Unlike in the medical scenario, Internet services have slightly different needs with respect to consent gathering:

a) If consent is to be gathered for a system or service under permanent evolution, changes in the service must be documented and reflected in the terms of use. This differs from one-time surgical operations in healthcare.
b) Hand-written signatures are rarely utilized as expression of consent in IT services. Hence, attribution, authenticity, and validity of an expression of consent must be collected using different techniques such as qualified electronic signatures [73]
c) In most scenarios, consent must be obtained before a user profile – and hence authentication password or token – even exists. Nevertheless, the association between user profile (i.e. personal data) and expression of consent (i.e. click on agreement button) must somehow be documented and stored in a tamper-proof way.

On the plus side, utilizing consent management systems reduces management efforts of the data controller and the processors. Once the system is up and running, the task of consent gathering is mostly automated, leaving the costly and scarce expert human resources free to focus on other tasks, such as determining whether a new consent must be gathered or not. A second clear advantage is the ability to integrate the consent management system with other management tools, such as CRM systems, audit and certification support, or legal affairs. Here, depending on the amount of integration, a huge potential for minimization of efforts exists, as the alternative would be to implement manual management procedures, binding costly human expertise.

On the downside, the efforts of implementing and integrating such a consent management system can be substantial. Depending on the degree of integration, the ramp-up of installing such a system may require extensive resources, and may pay off only partially later-on. The more integration is attempted, the higher the initial costs, but also the higher the resulting

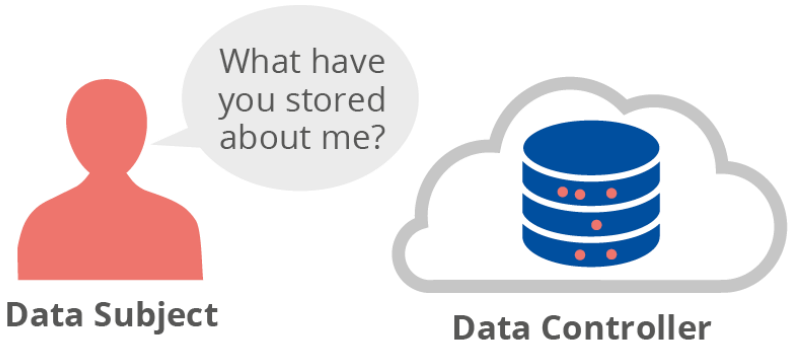
savings in the long term. This (common) imbalance may keep smaller enterprises from integrating such systems at all.

6.9 EXERCISING RIGHT OF ACCESS

As stipulated in Article 15 GDPR, every data controller has to provide all data subjects with access to the personal data stored about them in the data controller's systems, and to the extent in which data processors are participating in data processing. The same applies to data processors. However, the task of answering such an information requests on the right of access, especially in large, complex data processing networks with a multitude of data processors (and potentially additional joint data controllers, cf. GDPR Art. 26) can be very challenging. To address this challenge, many modern companies implement technical infrastructure and services for automating processing such access requests by providing e.g. privacy dashboards, as mentioned in 6.5.

Replying to a right of access request can be challenging in large, complex data processing networks with a multitude of data processors

Figure 3: Right of Access Request



Such a right of access service would provide an interface for data subjects whose data is processed by the organization in consideration. When triggered, the service automatically iterates over the data stores within the organization, collecting all personal data concerning the demanding individual, and delivering provides to the data subject the complete set of data collected this way towards the data subject. Ideally, the whole task is automated, so that no (or negligible) manual interaction at the side of the organization is required.

Having such an automated right of access service as part of an organization's internal or external management systems reduces greatly manual efforts when it comes to excessive amounts of right of access requests. While employees can easily come to their limits when demand grows up, technical infrastructure is usually easier to scale. Depending on the amount of such requests, such automated system may deliver significant savings to the organization.

At the same time connecting to the right of access service each new data storage, data sink or an additional data processor that gets an individual's data may also improve data management capabilities of the organization as a whole. Requests concerning data locations, data forwards, business partners involved in data processing, etc. can all be answered rather straight-forwardly just from the existing data flow infrastructures created and maintained for the right of access service.

On the downside, implementing such service requires additional processes to be defined, developed, and deployed along with the core functional services for data processing. This brings up a set of additional challenges to consider:

- **Authorization:** A human individual is only allowed to see and investigate its own personal data, not that of other data subjects. Hence, there must be some (technical

and/or organizational) means of authentication in place, to verify the authorization of the demanding individual. This authorization must, of course, guarantee validity of the authorization, hence must rely on high-level security techniques to validate the identity of a human individual. This may involve e.g. passport validation, two-factor authentication, or other similar means.

- **Delegated Authorization:** Sometimes, delegation of the right of access is possible, e.g. for under age children, legal custodians, attorneys, etc. In such cases, the authorization of the right of access request must be validated not just by verifying the identity of the demanding individual, but also by means of verifying the legal grounds for the transfer of authorization. Depending on the type of delegation of rights, this task may become arbitrarily complex (see also below).

- **Risk of Data Breach:** Disclosing the full set of personal data of one data subject to another data subject without valid authorization is equivalent to a severe data breach, which itself manifests a violation of the GDPR. At the same time, there may be a substantial interest in such right of access services by other actors than the concerned data subject, such as hackers, media, law enforcement, or relatives. Hence, the security risk for operating such a service is not negligible.

- **Completeness:** As was evident from the Schrems court cases, achieving completeness of the data disclosed in response to a right of access request is challenging. A recent study showed that only about 10% of companies provided a complete stack of customer data when demanded under Art. 15 GDPR [74]. However, an incomplete response to a right of access is a violation of Art. 15 GDPR, and would hence render the utilization of such right of access service ineffective. The main challenge here is how to identify all the data that belongs to a certain data subject in the huge set of data stores typically found at large data controller or data processor organizations. Sometimes, this task maps to querying databases with the customer identifier of the data subject (if available), but it may also include skimming through vast amounts of archived data, file systems, backups, derived data sets, or other type of information that is no longer directly and easily accessible or properly linked to the data subject's customer identifier – if such identifier exists at all.

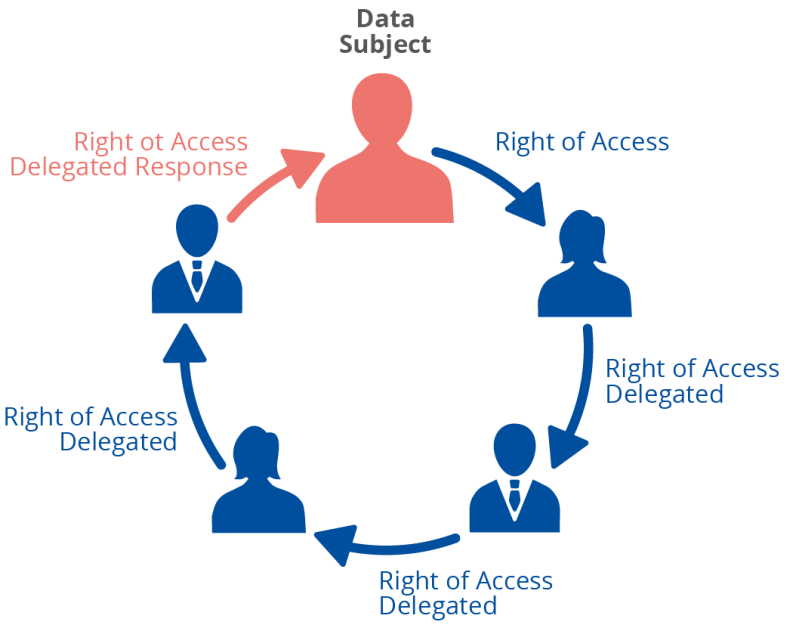
- **Correctness:** Similar to completeness, the data disclosed to the data subject must be correct, hence many not contain abbreviations, aggregations, internal censorship, or other access-blocking means. Also, its integrity must be maintained when delivered to the requesting individual. Hence, the implementation of such a right of access service must utilize sound technical means to guarantee correctness and integrity of the data contained in the response to the right of access demand.

- **Volume:** Personal profiles of active data subjects typically grow in size over the time of utilization of a service. Hence, the size of the response to a right of access request can easily grow into huge amounts of data. This causes a technical challenge of delivering the data to the requesting individual by reasonable means. E.g. the maximum size of allowed e-mail attachment can be easily reached, rendering an information mail as response to a right of access request infeasible. Print-outs are not just environmentally problematic but also do not fulfil the common requirements concerning right of access responses nor the demand for data portability as defined in Art. 20 GDPR. Common solutions consist in web driven downloads of compressed data archives via HTTP(S) or FTP(S), which also have some technical challenges for low-bandwidth areas.

6.9.1 Delegation of Access Rights Requests

Beyond the ability to delegate the authorization for performing a right of access request e.g. to a data custodian, the delegation of a right of access request may also include iteration over the full set of data controllers and data processors involved in a data processing activity [75]. In such cases, the delegation of the right of access is not only transferred to a dedicated single entity, but instead is merely passed along with the request itself to all sub-processors involved. More precisely, once the data subject demands an information according to the right of access, the data controller or data processor in charge contacts all of its sub-processors involved in this particular processing, and forwards the right of access request to each one of them. Those again contact their sub-processors and so forth, until the whole tree of data processors (and joint controllers) involved has been contacted. Each such delegated request is then answered with all information received from the sub-processors plus the information concerning the organization that was requested to provide data. Hence, a single request placed by the demanding individual is answered once, with a full set of information recursively collected from the whole set of sub-processors involved.

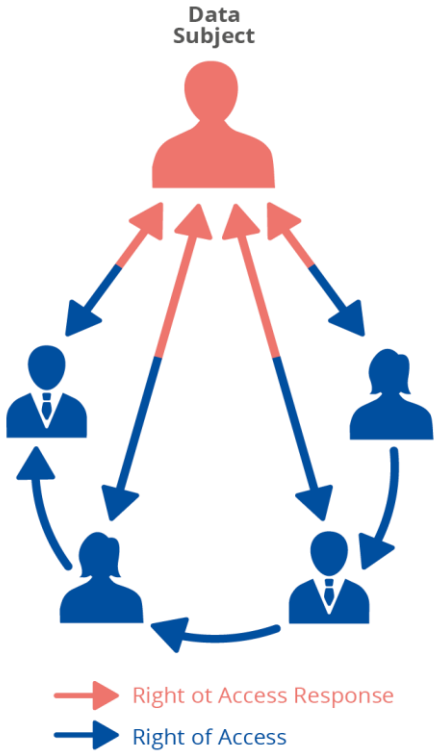
Figure 4: Recursive Right of Access Delegation



For the data subject, the advantage of such an infrastructure is that a single right of access request suffices to get a full view on the whole data processing activity, across all sub-processor borders. For data processors, the advantage of such data request infrastructure is its compositionality: details on the exact network of sub-processors, suppliers, service providers etc. can be easily hidden or masked in the single response sent back to the requesting predecessor in the processing tree. This way, the exact identity of an organization's business partners may be hidden from the previous processors, if considered a business secret. Nevertheless, the data collected for the original right of access request still is complete and sufficient for the data subject's needs.

The obvious drawback of this approach is again its implementation efforts and complexity: the delegated right of access services must be implemented and operated. Any requests related to the recursive nature may require more computational resources and longer execution times than a normal right of access response.

Figure 5: Iterative Right of Access Delegation



Alternatively, a data custodian organization can itself also provide the service of data collection to its data subjects. Unlike the recursive approach, in this case, the task of individually identifying and demanding right of access responses from all sub-processors in the processing network is performed iteratively by the data custodian, on behalf and by request of the data subject. Once the collection is completed, the resulting aggregated right of access response is then returned back to the demanding data subject. In this case, the advantage of getting a full picture on the processing at the data subject remains evident, whereas the data controllers and processors loose some control over what exactly is contained in such an aggregated right of access request.

6.10 EXERCISING RIGHT TO ERASURE, RIGHT TO RECTIFICATION

Similar to the right of access, the other data subjects' rights to erasure, rectification, blocking, restriction of processing, etc. can also be implemented in an equivalent way as dedicated services. Here, the infrastructure for the right of access services turns out to be very helpful, as it allows to easily identify all the data stores affected by the particular request. Also, a notification that a demand for the right to erasure or rectification was triggered can be sent to the data processors (or data controllers) responsible for this data.

7. CONCLUSIONS

Data protection principles, as set out in Article 5 of the GDPR and elaborated in terms of measures and safeguards in Article 25, are the goals that should be achieved when considering the design, implementation and deployment of a processing operation. From the technical side, the challenge is to translate these principles into tangible requirements and specifications by requirements by selecting, implementing and configuring appropriate technical and organizational measures and techniques over the complete lifecycle of the envisaged data processing. Engineering Data Protection into practise is not that straightforward though; depending on the level of risk, the context of the processing operation, the purposes of processing, the types, scope and volumes of personal data, the means and scale for processing, the state of the art, the cost, the translation into actionable requirements calls for a multidisciplinary approach. In addition, the evolving technological landscape and emerging technologies should also be taken into account as new challenges emerge such as lack of control and transparency, possible reusability or purpose "creep" with use of data, data inference and re-identification, profiling and automated decision making. The implementation of data protection principles in such contexts is challenging as they cannot be implemented in the traditional, "intuitive" way. Appropriate safeguards, both technical and organizational, must be integrated into the processing from the very early steps, as dictated by the Data protection by design obligation, and indeed the design process and related decision making needs to be underpinned by this obligation also.

This report attempted to provide a short overview of existing (security) technologies and techniques that can support the fulfilment of data protection principles and discuss possible strengths and their possible applicability in different processing operations. The remaining of this section presents the main conclusions to this end, together with specific recommendations for relevant stakeholders.

7.1 DEFINING THE MOST APPLICABLE TECHNIQUE

As it has been stressed also in past ENISA's reports, a number of technologies and techniques already exists but it is not straightforward for data controllers and data processors which one is applicable and better suited for each processing operation and for each context. More importantly it is not clear of how each technique should be engineered seamlessly into the processing operation practise in order to truly unfold their potentials and support achieving data protection principles.

The research community should continue exploring the deployment of (security) techniques and technologies that can support the practical implementation of data protection principles, with the support of the EU institutions in terms of policy guidance and research funding.

Regulators (e.g. Data Protection Authorities and the European Data Protection Board), the European Commission and the relevant EU institutions should disseminate the benefits of such technologies and techniques and provide guidance on their applicability and deployment.

Initiatives aimed to support engineers, such as the Internet Privacy Engineering Network (IPEN)[24], should be further supported by practitioners, researchers and academia.

[24] https://edps.europa.eu/data-protection/ipen-internet-privacy-engineering-network_en

7.2 ESTABLISHING THE STATE-OF-THE-ART

Proper implementation and engineering of discussed technologies and techniques is highly dependent on the state-of-the-art and the way that they are known and/or available to controllers. While not all techniques are equally effective, there might be certain implementation challenges or limitations with regard to each one. This is not only relevant to the choice of the technique itself, but also to the overall design of the processing operation.

Regulators (e.g. Data Protection Authorities and the European Data Protection Board) should discuss and promote good practices across the EU in relation to state-of-the-art solutions of relevant technologies and techniques. EU Institutions could promote such good practices by relevant publicly available documents.

7.3 DEMONSTRATE COMPLIANCE AND PROVIDE ASSURANCE

Further to guidance, data controllers and processors should be able to have some assurance on the soundness and correctness of their deployed processing operations while at the same time meeting their regulatory obligation to be able to demonstrate the overall level of protection they offer. Towards this direction, the provisions under Article 42 of the GDPR on data protection certification mechanisms, seals or marks could become useful tools not only to demonstrate compliance (and effectiveness) but also act as guidance when engineering data protection for processing operations. This is even more evident in emerging technologies, such as Artificial Intelligence, where the threat landscape, technological approaches, implementations and data protection challenges are evolving.

Regulators (e.g. Data Protection Authorities and the European Data Protection Board) and the European Commission should promote the establishment of relevant certification schemes, under Article 42 GDPR, to ensure proper engineering of data protection.

Regulators (e.g. Data Protection Authorities and the European Data Protection Board) should ensure that regulatory approaches, e.g. as regards new technologies and application sectors, take into account all possible entities and roles from the standpoint of data protection, while remaining technologically neutral.

8. REFERENCES

[1] ENISA, "Privacy by design in big data," 2015.

[2] EDPB, "Guidelines 4/2019 on Article 25 Data Protection by Design and by Default," 2019.

[3] ENISA, "Privacy and Data Protection by Design," 2015.

[4] M. Hansen, M. Jensen and M. Rost, "Protection Goals for Privacy Engineering," in *2015 IEEE Security and Privacy Workshops*, 2015.

[5] M. Colesky, J. H. Hoepman and C. Hillen, "A Critical Analysis of Privacy Design Strategies," in *2016 IEEE Security and Privacy Workshops (SPW)*, 2016.

[6] European Data Protection Supervisor, "Opinion 5/2018 Preliminary Opinion on privacy by design," 2018.

[7] ENISA, "Readiness Analysis for the Adoption and Evolution of Privacy Enhancing Technologies," 2016.

[8] ENISA, "PETs controls matrix - A systematic approach for assessing online and mobile privacy tools," 2016.

[9] Agencia Española de Protección de Datos (AEPD), "A Guide to Privacy by Design," 2019.

[10] ARTICLE 29 Data Protection Wokring Party, "Opinion 05/2014 on Anonymisation Techniques," 2014.

[11] ENISA, "Recommendations on shaping technology according to GDPR provisions - An overview on data pseudonymisation," 2019.

[12] ENISA, "Pseudonymisation techniques and best practices".

[13] ENISA, "Data Pseudonymisation: Advanced Techniques and Use Cases," 2021.

[14] F. Bonchi, L. V. Lakshmanan and H. Wang, "Trajectory anonymity in publishing personal mobility data," *SIGKDD Explor,* vol. 31, no. 1, pp. 30-42, 2011.

[15] O. Abul, F. Bonchi and M. Nanni, "Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases," in *IEEE 24th International Conference on Data Engineering (ICDE 08)*, 2008.

[16] O. Abul, F. Bonchi and M. Nanni, "Anonymization of moving objects databases by clustering and perturbation," *Information Systems,* vol. 35, no. 8, pp. 849-910, 2010.

[17] R. Chen, G. Acs and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," in *2012 ACM conference on Computer and communications security*, 2012.

[18] Datalisynet, "A guide to the anonymisation of personal data," 2015.

[19] ICO, "Introduction to anonymisation," 2021.

[20] S. L. Garfinkel, "NISTIR 8053 De-Identification of Personal Information," 2015.

[21] B. Fung, K. Wang, R. Chen and P. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys,* vol. 42, no. 4, pp. 1-53, 2010.

[22] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 10, no. 5, p. 557–570, 2002.

[23] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, 2006.

[24] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, 2007.

[25] S. R. Ganta, S. P. Kasiviswanathan and A. Smith, "Composition attacks and auxiliary information in data privacy," in *14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.

[26] A. Meyerson and R. Williams, "On the complexity of optimal K-anonymity," in *23rd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2004.

[27] K. E. Emam and F. K. Dankar, "Protecting privacy using k-anonymity," *Journal of the American Medical Informatics Association,* vol. 15, no. 5, p. 627–637.

[28] C. Dwork and A. Roth, "The Algorithmic Foundationsof Differential Privacy," Foundations and Trends in Theoretical Computer Science, 2014.

[29] C. Dwork, N. Kohli and D. Mulligan, "Differential Privacy in Practice: Expose your Epsilons!," *Journal of Privacy and Confidentiality,* vol. 9, no. 2, 2019.

[30] C. Dwork, "Differential Privacy," in *International Colloquium on Automata, Languages, and Programming (ICALP 2006)*, 2006.

[31] G. Ács and C. Castelluccia, "I Have a DREAM! (DiffeRentially privatE smArt Metering)," in *International Workshop on Information Hiding (IH 2011)*, 2011.

[32] C. Clifton and T. Tassa, "On syntactic anonymity and differential privacy," in *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 2013.

[33] G. Acs and C. Castelluccia, "A case study: privacy preserving release of spatio-temporal density in Paris," in *20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

[34] M. A.Will and R. Ko, "A guide to homomorphic encryption," in *The Cloud Security Ecosystem*, Syngress, 2015, p. 101127.

[35] C. Gentry and S. Halevi, "Implementing Gentry's Fully-Homomorphic Encryption Scheme," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT 2011)*, 2011.

[36] L. Lamport, R. Shostak and M. Pease, "The Byzantine Generals Problem," *SRI International,* 1982.

[37] P. Bogetoft, D. L. Christensen, I. Damgård, M. Geisler, T. Jakobsen, M. Krøigaard, J. D. Nielsen, J. B. Nielsen, K. Nielsen, J. Pagter, M. Schwartzbach and T. Toft, "Secure Multiparty Computation Goes Live," in *Financial Cryptography and Data Security (FC 2009)*, 2009.

[38] D. Asonov, "Private Information Retrieval – An Overview and Current Trends," 2011.

[39] R. Ostrovsky and W. E. Skeith, "A Survey of Single-Database Private Information Retrieval: Techniques and Applications," in *PKC 2007: Public Key Cryptography*, 2007.

[40] W. Gasarch, "A Survey on Private Information Retrieval," *Bulletin of the EATCS,* vol. 82, pp. 72-107, 2004.

[41] K. Ermoshina, F. Musiani and H. Halpin, "End-to-End Encrypted Messaging Protocols: An Overview," in *International Conference on Internet Science (INSCI)*, 2016.

[42] N. Unger, S. Dechand, J. Bonneau, S. Fahl, H. Perl, I. Goldberg and M. Smith, "SoK: Secure Messaging," in *2015 IEEE Symposium on Security and Privacy*, 2015.

[43] European Data Protection Board, "Recommendations 01/2020 on measures that supplement transfer tools to ensure compliance with the EU level of protection of personal data," 2020.

[44] European Data Protection Board, "Statement of the EDPB on the revision of the ePrivacy Regulation and its impact on the protection of individuals with regard to the privacy and confidentiality of their communications," 2018.

[45] Y. Gilad, "Metadata-Private Communication for the 99%," *Communications of the ACM,* vol. 62, no. 9, pp. 86-93, 2019.

[46] M. Reed, P. Syverson and D. Goldschlag, "Anonymous connections and onion routing," *IEEE Journal on Selected Areas in Communications,* vol. 16, no. 4, pp. 482 - 494, 1998.

[47] M. Reed, P. Syverson and D. Goldschlag, "Anonymous connections and onion routing," *IEEE Journal on Selected Areas in Communications,* vol. 16, no. 4, pp. 482-494, 1998.

[48] ENISA, "Reinforcing trust and security in the area of electronic communications and online services: Sketching the notion of “state-of-the-art” for SMEs in security of personal data processing," 2019.

[49] J. Camenisch, A. Lehmann, G. Neven and A. Rial, "Privacy-Preserving Auditing for Attribute-Based Credentials," in *European Symposium on Research in Computer Security (ESORICS 2014)*, 2014.

[50] K. Rannenberg, J. Camenisch and A. Sabouri, Attribute-based Credentials for Trust, Springer, 2015.

[51] S. Goldwasser, S. Micali and C. Rackoff, "The knowledge complexity of interactive proof-systems," in *Seventeenth annual ACM symposium on Theory of Computing (STOC 85)*, 1985.

[52] J.-J. Quisquater, M. Quisquater, M. Quisquater, M. Quisquater, L. Guillou, M. A. Guillou, G. Guillou, A. Guillou, G. Guillou and S. Guillou, "How to Explain Zero-Knowledge Protocols to Your Children," in *Advances in Cryptology (CRYPTO 89)*, 1990.

[53] M. Blum, P. Feldman and S. Micali, "Non-interactive zero-knowledge and its applications," in *Twentieth annual ACM symposium on Theory of computing (STOC 88)*, 1988.

[54] ARTICLE 29 Data Protection Working Party, "Opinion 10/2004 on More Harmonised Information Provisions," 2004.

[55] ARTICLE 29 Data Protection Working Party, "Guidelines on transparency under Regulation 2016/679," 2018.

[56] Information Commissioner’s Office (ICO), "Age appropriate design: a code of practice for online services," [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/ico-codes-of-practice/age-appropriate-design-a-code-of-practice-for-online-services/.

[57] Forbrukerrådet (Norwegian Consumer Council), "Deceived by design," 2018.

[58] L. E.Holtz, K. Nocun and M. Hansen, "Towards Displaying Privacy Information with Icons," in *Privacy and Identity 2010: Privacy and Identity Management for Life*, 2010.

[59] L. Edwards and W. Abel, "The Use of Privacy Icons and Standard Contract Terms for Generating Consumer Trust and Confidence in Digital Services," 2014.

[60] P. Balboni and K. Francis, "Maastricht University Data Protection as a Corporate Social Responsibility (UM DPCSR) Research Project: UM DPCSR Icons Version 1.0," 2020.

[61] H. Habib, Y. Zou, Y. Yao, A. Acquisti, L. Cranor, J. Reidenberg, N. Sadeh and F. Schaub, "Toggles, Dollar Signs, and Triangles: How to (In)Effectively Convey Privacy Choices with Icons and Link Texts," in *2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[62] M. C.-M. S. Pearson, "Sticky Policies: An Approach for Managing Privacy across Multiple Parties," *Computer,* vol. 44, no. 9, pp. 60-68, 2011.

[63] M. Hils, D. W. Woods and R. Böhme, "Privacy Preference Signals: Past, Present and Future," *Proceedings on Privacy Enhancing Technologies,* vol. 2021, no. 4, pp. 249-269, 2021.

[64] P3P, "THE PLATFORM FOR PRIVACY PREFERENCES 1.1 (P3P1.1)," [Online]. Available: https://www.w3.org/standards/history/P3P11.

[65] W3C, "A P3P Preference Exchange Language 1.0 (APPEL1.0)," [Online]. Available: https://www.w3.org/TR/2002/WD-P3P-preferences-20020415/.

[66] European Commission, "Factsheet: Access to Base Registries in Estonia," 2017.

[67] S. Fischer-Hübner, J. Angulo, F. Karegar and T. Pulls, "Transparency, Privacy and Trust–Technology for Tracking and Controlling My Data Disclosures: Does This Work?," in *Trust Management X: 10th IFIP WG 11.11 International Conference, IFIPTM 2016*, 2016.

[68] B. W. Schermer, B. Custers and S. v. d. Hof, "The crisis of consent: how stronger legal protection may lead to weaker consent in data protection," *Ethics and Information Technology,* vol. 16, pp. 171-184, 2014.

[69] M. Hils, D. W. Woods and R. Böhme, "Measuring the Emergence of Consent Management on the Web," in *ACM Internet Measurement Conference*, 2020.

[70] C. Santos, M. Nouwens, M. Toth, N. Bielova and V. Roca, "Consent Management Platforms Under the GDPR: Processors and/or Controllers?," in *Annual Privacy Forum 2021*, 2021.

[71] M. R. Asghar, T. Lee, M. M. Baig, E. Ullah, G. Russello and G. Dobbie, "A Review of Privacy and Consent Management in Healthcare: A Focus on Emerging Data Sources," in *2017 IEEE 13th International Conference on e-Science (e-Science)*, 2017.

[72] M. Benchoufi, .. Porcher and P. Ravaud, "Blockchain protocols in clinical trials: Transparency and traceability of consent," F1000Research, 2017.

[73] ENISA, "Security guidelines on the appropriate use of qualified electronic signatures," 2017.

[74] J. Tolsdorf, M. Fischer and L. L. Iacono, "A Case Study on the Implementation of the Right of Access in Privacy Dashboards," in *Annual Privacy Forum 2021*, 2021.

[75] R. Herkenhöner, H. d. Meer, M. Jensen and H. C. Pöhls, "Towards Automated Processing of the Right of Access in Inter-organizational Web Service Compositions," in *2010 6th World Congress on Services*, 2010.

TP-01-22-045-EN-N

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, the European Union Agency for Cybersecurity contributes to EU cyber policy, enhances the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperates with Member States and EU bodies, and helps Europe prepare for the cyber challenges of tomorrow. Through knowledge sharing, capacity building and awareness raising, the Agency works together with its key stakeholders to strengthen trust in the connected economy, to boost resilience of the Union's infrastructure, and, ultimately, to keep Europe's society and citizens digitally secure. More information about ENISA and its work can be found here: www.enisa.europa.eu.

ENISA
European Union Agency for Cybersecurity

Athens Office
Agamemnonos 14, Chalandri 15231, Attiki, Greece

Heraklion Office
95 Nikolaou Plastira
700 13 Vassilika Vouton, Heraklion, Greece

enisa.europa.eu

ISBN: 978-92-9204-556-2
DOI: 10.2824/09079