Guideline

PSEUDONYMISATION, MINIMISATION AND ENCRYPTION

Document Code	**14e-HD/SG/HDCV/FSOFT**
Version	**1.4**
Effective date	**01-Aug-2023**

TABLE OF CONTENT

RECORD OF CHANGE

No	Effective Date	Version	Change Description	Reason	Reviewer	Final Reviewer	Approver
1	01-May-2021	1.0	Newly issued	Describe how Personal Data Protection can be pseudonymized, minimized and encrypted	TrangNN4	Michael Hering	HoanNK
2	01-Oct-2021	1.1	1.2 added: statement_PIMS scope_V1.0, 5.2 added: statement_PIMS scope_V1.0	Legal requirement	TrangNN4	Michael Hering	HoanNK
3	01-Apr-2022	1.2	1.2 added: Policy_PIMS scope_V1.1 5.2 13 added PIPL, 5.2 14 added: PDPL, UAR, Decree-Law No. 45 of 2021 5.2 16 added: Decree of the Vietnamese Government: Nghị Định Quy Định Về Bảo Vệ Dữ Liệu Cá Nhân 5.2 17 PDP_ Handbook_Version_V 3.2 5.2 18: 15e-HD/SG/HDCV/FSOFT	Biannually revision	LinhDTD1	Michael Hering	HoanNK
4	01-Nov-2022	1.3	Added 5.3 Data Protection Law, Vietnam, Overview. Added 5.2.15 Republic Act 10173 Data privacy Act 2012 Added 5.2 17 PDPA Added 5.2 18 TISAX	Biannually revision	LinhDTD1	Michael Hering	HoanNK
4	01-Aug-2023	1.4	Adjust document version numbers added 5.2 14, 18 changed 5.2 22: Came in force 07/2023 changed 5.3 PDPD was finalized and was coming in force 07/2023	Biannually revision	LinhDTD1	Michael Hering	HoanNK

1 INTRODUCTION

FPT Software Company, Ltd. ("FPT Software" hereinafter) Corporate Data Protection Policy, guidelines, procedures and templates lay out strict requirements for processing personal data pertaining to customers, business partners, employees or any other individual. It meets the requirements of the European Data Protection Regulation/Directive as well as other national Data Protection Regulations and ensures compliance with the principles of national and international data protection laws in force all over the world. The policy, guidelines, procedures and templates set a globally applicable data protection and security standard for FPT Software and regulates the sharing of information between FPT Software, subsidiaries, and legal entities. FPT Software have established guiding data protection principles – among them transparency, data economy and data security – as FPT Software Personal Data Protection Handbook and ISM guidelines.

1.1 Purpose

This guideline describes how Personal Data Protection can be pseudonymized, minimized and encrypted. Pseudonymizing personal data reduces the risks to data subjects and helps FPT Software as a controller or processor to meet its data protection obligations by ensuring that the additional information that attributes personal data to a specific data subject is kept separately.

The GDPR defined pseudonymization as “the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.”.

Pseudonymization is a well-known de-identification process that has gained additional attention following the adoption of GDPR and other national data protections laws or regulations, where it is referenced as both a security and data protection by design mechanism. In addition, in the GDPR context, pseudonymization can motivate the relaxation, to a certain degree, of data FPT Software’ legal obligations if properly applied.

Starting from the definition of pseudonymization (as well as its differences from other technologies, such as anonymization and encryption), this guideline first discusses the core data protection benefits of pseudonymization. It presents techniques that are utilized for pseudonymization, such as hashing, hashing with key or salt, encryption, tokenization, as well as other relevant approaches.

1.2 Application Scope

See Policy_PIMS scope_V1.3.

This guideline is binding for all departments and functions globally which are involved in personal identifiable information processing.

1.3 Application of national Laws

The Data Protection Policy, guidelines and templates comprises the internationally accepted data privacy principles without replacing the existing national laws. It supplements the national data privacy laws. The relevant national law will take precedence in the event that it conflicts with the Data Protection Policy and guidelines, or it has stricter requirements than this Policy and guidelines. The content of the Data Protection Policy and guidelines must also be observed in the absence of corresponding national legislation. The reporting requirements for data processing under national laws must be observed.

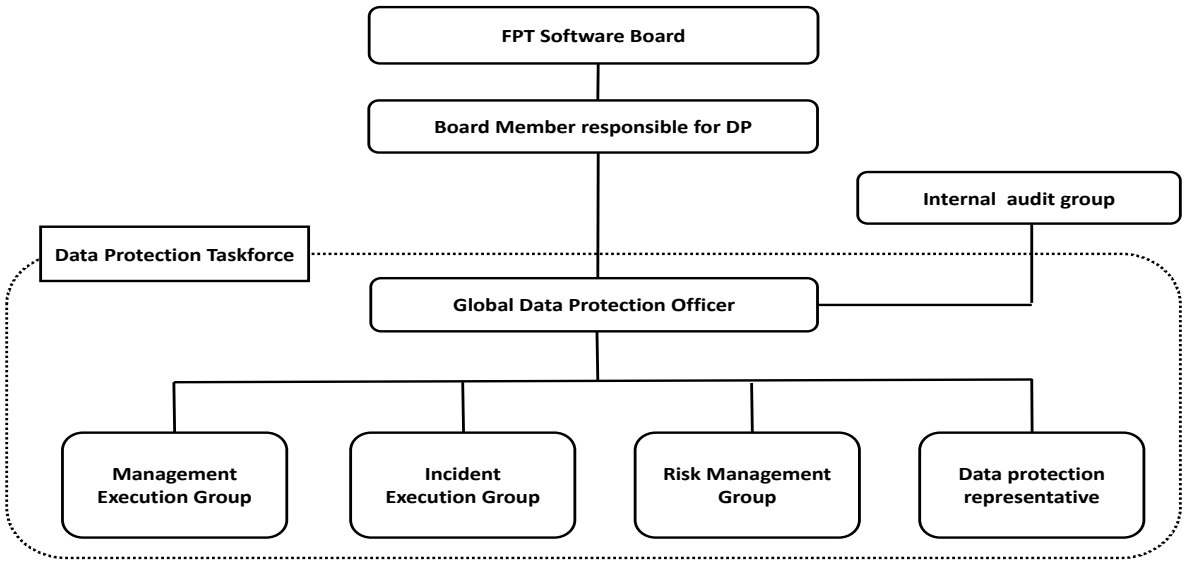
Each subsidiary or legal entity of FPT Software is responsible for compliance with the Data Protection Policy, this guideline and the legal obligations. If there is reason to believe that legal obligations contradict the duties under the Data Protection Policy or the guidelines, the relevant subsidiary or legal entity must inform the Global Data Protection Officer. In the event of conflicts between national legislation, the Data Protection Policy, and this guideline, FPT Software will work with the relevant subsidiary or legal entity of FPT Software to find a practical solution that meets the purpose of the Data Protection Policy and this guideline.

1.4 Responsibility

The Global Data Protection Officer, appointed by the FPT Software Board Member responsible for Data Protection on behalf of the CEO of FPT Software is fully responsible.

The Global Data Protection Officer (GDPO) is an enterprise security leadership role required by the General Data Protection Regulation (GDPR) and other national laws. The GDPO is responsible for overseeing data protection strategy and implementation to ensure compliance with GDPR requirements and other national Personal Data Protection Acts. The primary role of the GDPO is to ensure that organization processes, the personal data of employees, customers, providers or any other individuals are in compliance with the applicable data protection rules. GDPO should be able to perform the duties independently.

GDPO is responsible for recommendation and observation of the exercise on pseudonymization, minimization and encryption. GDPO must ensure that all departments of the company are following the company guidelines and the respective laws.



2 GUIDELINE CONTENT

2.1 Minimization

Data minimization applies to the third principle of data protection introduced by the Data Protection Directive 95/46/EC and has been incorporated into the GDPR.

The third principle of data protection specifies that personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

This obliges FPT Software to obtain and use only those pieces of information that are necessary for the FPT Software's purpose(s) for processing such information. Holding any additional personal data on individuals is unlawful.

2.2 Encryption

Every mobile device of FPT Software must have encrypted storage (hard disk, SSD….)
By their very nature mobile devices such as laptops, smartphones and tablets have a high risk of loss or theft. Encryption of the data contained on the device provides an assurance that, if this happens, the risk of unauthorized or unlawful access is significantly minimized.
Sensitive and personal information that is on a laptop must, clearly, be encrypted. Of course, any other confidential information – financial data, customer information, and so on – should also be encrypted to protect it if the laptop is ever lost or stolen. The drawback with encryption solutions that only encrypt those files that contain confidential information is that laptop users don't always ensure they always save data into these folders, and these encryption solutions do not automatically encrypt temporary files or caches. FPT Software mobile devices must have a whole-disk encryption. This solution will automatically encrypt any portable storage media – such as USB sticks and CD-ROMs – to which encrypted data might be exported.
The configuration requirements for the encryption solution are be documented, and the maintenance of this configuration standard is subject to regular monitoring and technical checking.

Recommendation of cryptographic algorithms			
Classification	**USA (NIST)**	**Europe (ECRYPT)**	**Korea (KISA)**
Symmetric key encryption algorithm	AES-128/192/256 3TDEA	AES-128/192/256 Blowfish KASUMI 3TDEA	SEED, HIGHT ARIA-128/192/256
Public key cryptography algorithm	RSA- 2048	RSAES-OAEP RSAES-PKCS1	RSAES-OAEP
One-way hash algorithm	SHA-224/256/384/512	SHA-224/256/384/512 Whirlpool	SHA-224/256/384/512

2.3 Pseudonymization

Pseudonymization is a well-known de-identification process that has gained additional attention following the adoption of GDPR, where it is referenced as both a security and data protection by design mechanism. In addition, in the GDPR context, pseudonymization can motivate the relaxation, to a certain degree, of data controllers' legal obligations if properly applied.

It is highly important for both data controllers/data processor and data subjects.

2.4 Relevant Pseudonymization Scenarios

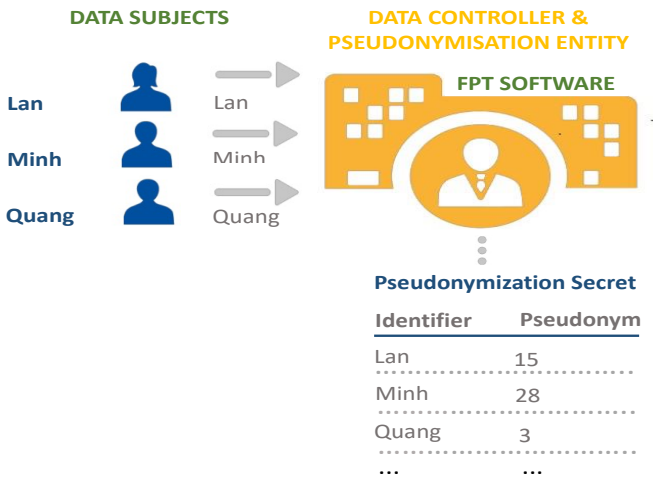
Pseudonymization has an important role in GDPR as a security measure (art. 32 GDPR), as well as in the context of data protection by design (art. 25 GDPR).

The benefit of pseudonymization is to hide the identity of the data subjects from any third party (i.e., other than the pseudonymization entity) in the context of a specific data processing operation. Pseudonymization can go beyond hiding real identities into supporting the data protection goal of unlink ability, i.e., reducing the risk that privacy-relevant data can be linked across different data processing domains. Furthermore, pseudonymization (being itself a data minimization measure) can contribute towards the principle of data minimization under GDPR, for example in cases where the controller does not need to have access to the real identities of data subjects but only to their pseudonyms.

First Scenario: Pseudonymization internally

Data are collected directly from the data subjects and pseudonymized by the data controller, for subsequent internal processing.

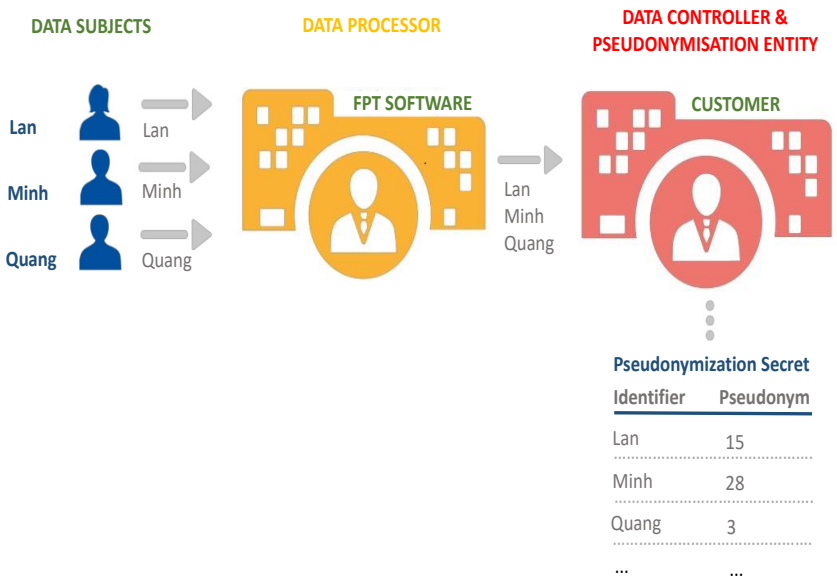
The data controller/FPT Software has the role of the pseudonymization entity, as it performs the selection and assignment of pseudonyms to identifiers. It must be pointed out that the data subjects do not necessarily know their particular pseudonym, as the pseudonymization secret (e.g., the pseudonymization mapping table in this example), is known only to FPT Software. The role of pseudonymization in this case is to enhance the security of personal data either for internal use (e.g., sharing between different legal entities or subsidiaries of the FPT Software) or in the case of a data protection incident.



Second Scenario: Processor involved in pseudonymization

A data processor is involved in the process by obtaining the identifiers from the data subjects (on behalf of the controller). The pseudonymization is still performed by the controller.

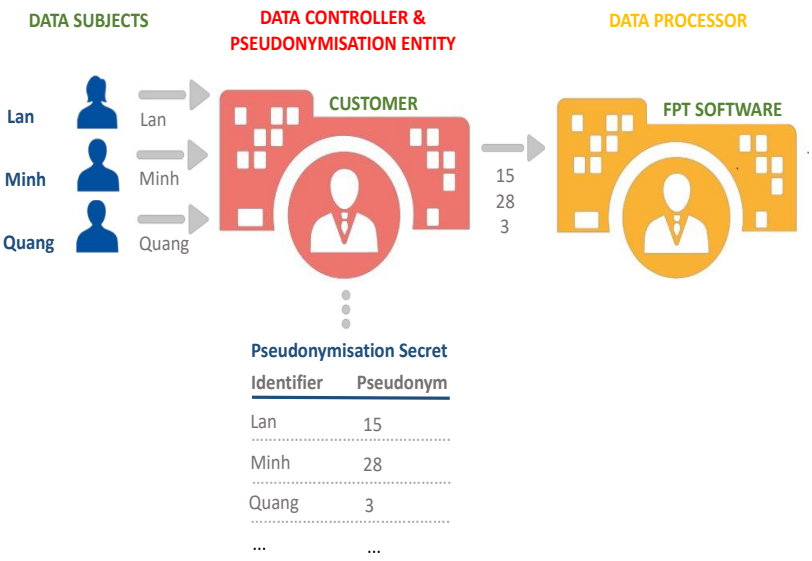
A dedicated data processor/FPT Software is given the task to collect the identifiers from the data subjects and forward this information to a subsequent data controller (FPT Software customer), which finally performs the pseudonymization. The controller is the pseudonymization entity. An example for such a scenario might be a BPO service provider offering data collection services on behalf of the data controller. Then, the controller still is in charge of applying data pseudonymization prior to any subsequent processing. The goals for pseudonymization are still the same as in scenario 1.



Third Scenario: Pseudonymized data (by customer) processed by a processor

The data controller (customer) performs the pseudonymization, but the processor (FPT Software) is not involved in the process only receives the pseudonymized data from the controller.

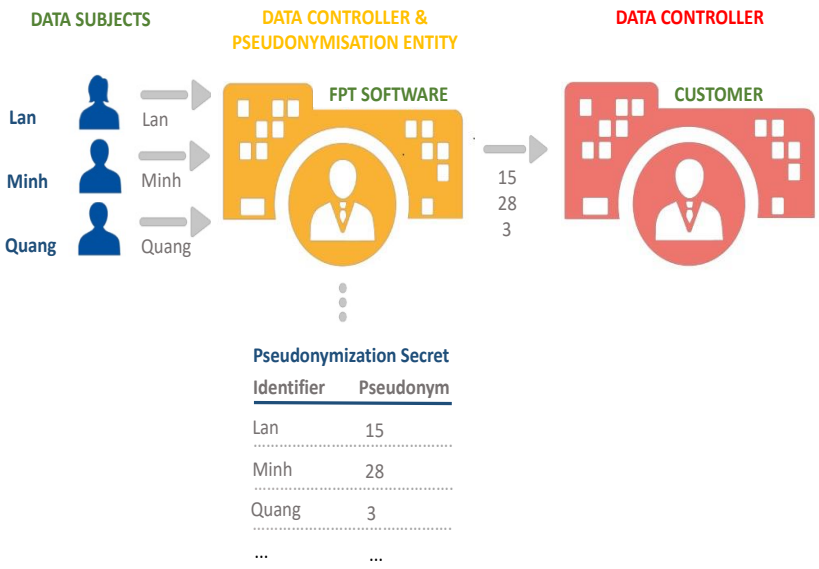
A data controller (customer of FPT Software) is collecting data and performing the task of data pseudonymization (role: pseudonymization entity). The difference with scenarios 2 is that now this data controller forwards the pseudonymized data to a subsequent data processor (Beta Inc.), e.g. for statistical analysis, or persistent data storage. In this scenario, the protection goal provided by data pseudonymization can unfold: Beta Inc. does not learn the identifiers of the data subjects, thus is not directly able to re-identify the natural persons behind the data (assuming that no other attribute that could lead to re-identification is available to Beta Inc.). In this way, pseudonymization protects the security of the data with regard to the processor.



Fourth Scenario: Processor as pseudonymization entity

In this case the task of pseudonymization is assigned by the controller to a data processor (e.g., managed service provider that manages the pseudonymization secret)

Personal data are sent by the data subjects to a data processor (FPT Software), which subsequently performs the pseudonymization, thus acting as the pseudonymization entity on behalf of the controller (customer). The pseudonymized data is then forwarded to the data controller. In this scenario, only the pseudonymized data are stored on the controller's side. In this way, security at controller's level is enhanced through data de-identification (e.g., in case of data breach at controller's side). Still, in all cases the controller is able to re-identify the data subjects through the data processor. Moreover, security at processor's side becomes of significant importance.



3 PSEUDONYMIZATION TECHNIQUES

In principle, a pseudonymization function maps identifiers to pseudonyms. There is one fundamental requirement for a pseudonymization function. Let us consider two different identifiers $Id1$ and $Id2$ and their corresponding pseudonyms $pseudo1$ and $pseudo2$. A pseudonymization function must verify that $pseudo1$ is different than $pseudo2$. Otherwise, the recovery of the identifier could be ambiguous: the pseudonymization entity cannot determine if $pseudo1$ corresponds to $Id1$ or $Id2$. A single identifier Id can be associated to multiple pseudonyms ($pseudo1$, $pseudo2$...) as long as it is possible for the pseudonymization entity to invert this operation. In all cases, according to the definition of pseudonymization, there exists some additional information that allows the association of the pseudonyms with the original identifiers; this is the pseudonymization secret. The simplest case of pseudonymization secret is the pseudonymization mapping table.

In the following chapters, the main options available to pseudonymize a single identifier are first defined. The different policies available for pseudonymization are then described, comparing their implementation characteristics. A reference to the main criteria that a controller may use to select a pseudonymization technique is also made. Possibilities of recovery of pseudonymization by the pseudonymization entity are described.

3.1 Single identifier pseudonymization

Possible approaches for a pseudonymization of a single identifier, together with relevant advantages and constraints are explained in the following paragraphs.

Counter

Counter is the simplest pseudonymization function. The identifiers are substituted by a number chosen by a monotonic counter. First, a seed s is set to 0 (for instance) and then it is incremented. It is critical that the values produced by the counter never repeat to prevent any ambiguity.

The advantages of the counter rest with its simplicity, which make it suitable for small and not complex datasets. In terms of data protection, the counter provides for pseudonyms with no connection to the initial identifiers (although the sequential character of the counter can still provide information on the order of the data within a dataset). This solution may have implementation and scalability issues in cases of large and more sophisticated datasets.

Random number generator (RNG)

RNG is a mechanism that produces values in a set that have an equal probability of being selected from the total population of possibilities and, hence, are unpredictable. This approach is similar to the counter with the difference that a random number is assigned to the identifier.

There are two options are to create this mapping: a true random number generator or a cryptographic pseudo-random generator. It should be noted that in both cases, without due care, collisions can occur. A collision is the case of two identifiers being associated to the same pseudonym. The probability that a collision will appear is related to the well-known birthday paradox.

RNG provides strong data protection (as, contrary to the counter, a random number is used to create each pseudonym, thus it is difficult to extract information regarding the initial identifier, unless the mapping table is compromised). Collisions may be an issue, as well as scalability (the complete pseudonymization mapping table must be stored), depending on the implementation scenario.

Cryptographic hash function

A cryptographic hash function takes input strings of arbitrary length and maps them to fixed length outputs. It satisfies the following properties:

One-way: it is computationally infeasible to find any input that maps to any pre-specified output.

Collision free: it is computationally infeasible to find any two distinct inputs that map to the same output.

A cryptographic hash function is directly applied to the identifier to obtain the corresponding pseudonym: $Pseudo = H(Id)$. The domain of the pseudonym depends on the length of the digest produced by the function.

While a hash function can significantly contribute towards data integrity, it is generally considered weak as a pseudonymization technique as it is prone to brute force and dictionary attacks.

Message authentication code (MAC)

MAC can be seen as a keyed-hash function. It is very similar to the previous solution except that a secret key is introduced to generate the pseudonym. Without the knowledge of this key, it is not possible to map the identifiers and the pseudonyms. HMAC is by far the most popular design of message authentication code used in Internet protocols.

MAC is generally considered as a robust pseudonymization technique from a data protection point of view, since reverting the pseudonym is infeasible, as long as the key has not been compromised. Different variations of the method may apply with different utility and scalability requirements of the pseudonymization entity.

Encryption

In scope are mainly symmetric (deterministic) encryption and in particular block ciphers like the Advanced Encryption Standard (AES) and their modes of operation. The block cipher is used to encrypt an identifier using a secret key, which is both the pseudonymization secret and the recovery secret. Using block ciphers for pseudonymization requires to deal with the block size. The size of the identifiers can be smaller or larger than the input block size of block cipher. If the identifiers' size is smaller, padding must be considered. In the case where the identifiers' size is larger than the block size, there are two options that can be used to solve this problem; the identifiers can be compressed into something smaller than the block size; if compression is not an option available, a mode of operation (like the counter mode CTR) can be used. This last option requires managing an extra parameter, the initialization vector.

Encryption is a robust pseudonymization technique, with several properties similar to MAC. Even focusing on deterministic encryption schemes, probabilistic encryption is an alternative, which could be

used especially in cases where there is need to derive different pseudonyms for the same identifier (like fully randomized pseudonymization policy).

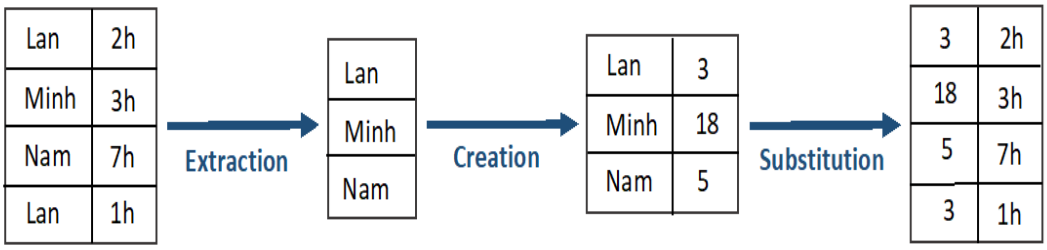
3.2 Pseudonymization policies

The pseudonymization technique is essential, the policy (or mode) of implementation of pseudonymization is equally important to its practical application.

This chapter considers the more general problem of the pseudonymization of a database or any document which contains k identifiers. Example: An identifier Id which appears several times in two datasets A and B. After pseudonymization, the identifier Id is substituted with respect to one of the following policies: deterministic pseudonymization, document-randomized pseudonymization and fully randomized pseudonymization.

Deterministic pseudonymization

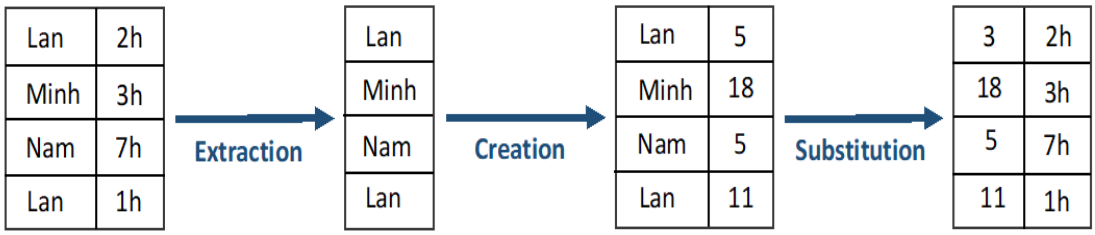
In all the databases and each time, it appears, Id is always replaced by the same pseudonym $pseudo$. It is consistent within a database and between different databases. The first step to implement this policy is to extract the list of unique identifiers contained in the database. Then, this list is mapped to the pseudonyms and finally the identifiers are substituted to the pseudonyms in the database.



All pseudonymization techniques mentioned before can be directly used to implement deterministic pseudonymization.

Document-randomized pseudonymization

Each time Id appears in a database, it is substituted with a different pseudonym ($pseudo1$, $pseudo2$,..). However, Id is always mapped to the same collection of ($pseudo1$, $pseudo2$) in the dataset A and B.



The pseudonymization is only consistent between different databases in this case. The mapping table is created this time using all the identifiers contained in the database. Each occurrence of a given identifier is treated independently.

Fully randomized pseudonymization

For any occurrences of Id within a database A or B, Id is replaced by a different pseudonym ($pseudo1$, $pseudo2$). This case is fully randomized pseudonymization**.** This policy can be viewed as a further extension of document randomized pseudonymization. In fact, the two policies have the same behavior when they are applied on a single document. If the same document is pseudonymized twice with fully randomized pseudonymization, two different outputs are obtained. With document-randomized pseudonymization, the same output would have been obtained twice. Means in document-randomized pseudonymization the randomness is selective, whereas in fully randomized pseudonymization randomness is global (it applies to any record).

3.3 Choosing a pseudonymization technique and policy

The choice of a pseudonymization technique and policy depends on different parameters, primarily the data protection level and the utility of the pseudonymized dataset. In terms of protection, RNG, message authentication codes and encryption are stronger techniques as they thwart by design exhaustive search, dictionary search and guesswork. Utility requirements might lead the pseudonymization entity towards a combination of different approaches or variations of a selected approach. Similarly, with regard to pseudonymization policies, fully- randomized pseudonymization offers the best protection level but prevents any comparison between databases. Document-randomized and deterministic functions provide utility but allow likability between records

The pseudonymization entity may be concerned by the complexity associated to a certain scheme in terms of implementation and scalability. How simple is it to apply pseudonymization to the identifiers and does pseudonymization has any impact on the database size?

Method	Identifier size	Pseudonym size m in bits
Counter	Any	$m = log_2 k$
Random Number Generator	Any	$m >> 2log_2 k$
Hash function	Any	Fixed or $m >> 2log_2 k$
Message Auth. Codes	Any	Fixed or $m >> 2log_2 k$
Encryption	Fixed[20]	Fixed or same as indentifier

Most solutions can be applied on identifiers of variable size except for certain choices in the case of encryption. The size of the pseudonym depends on k, the number of the identifiers contained in the database. For random number generator, hash function and message authentication code, there is a probability of collision: the size of the pseudonym must be chosen carefully (see birthday paradox). Hash functions and message authentication codes are suitably designed so as to ensure that the digest size prevents any risks of collision. The size of the pseudonyms produced by an encryption scheme can be fixed or equal to the size of the original identifier. Sheet before presents the scalability of the aforementioned approaches with regards to the recovery function.

3.4 Recovery

By definition, the use of additional information is central to pseudonymization, the pseudonymization entity must implement a recovery mechanism. This mechanism can be more or less complex depending on the pseudonymization function. In general, they consist of the use of a pseudonym $pseudo$ and a pseudonymization secret S to recover the corresponding identifier Id. This case can occur for example when the pseudonymization entity has detected an anomaly in its system and needs to contact the designated entities. The "anomaly" can be for instance a data breach and the pseudonymization entity needs to notify the data subjects under GDPR. In addition, the recovery mechanism might be necessary in order to allow for the exercise of data subjects rights (articles 12-21 GDPR).

Method	Recovery based on pseudonym
Counter	Mapping table
Random Number Generator	Mapping table
Hash function	Mapping table
Message Auth. Codes	Mapping table
Encryption	Decryption

Most methods require the pseudonymization entity to keep the mapping table between the identifiers and the pseudonyms to perform identifier recovery with the exception of encryption. Indeed, decryption can be directly applied on the identifier.

3.5 Protection of the pseudonymization secret

In order for pseudonymization to be efficient, the pseudonymization entity must always protect the pseudonymization secret by proper technical and organizational measures. This clearly depends on the specific pseudonymization scenario.

Firstly, the pseudonymization secret must be isolated from the dataset, i.e., the pseudonymization secret and the dataset must never be handled in the same file. Secondly, the pseudonymization secret must be securely deleted from any insecure media (memory storage and systems). Thirdly, strong access control policies must ensure that only authorized entities have access to this secret. A secure logging system must keep track of all the access requests made to the secret. Finally, the pseudonymization secret must be encrypted if it is stored on a computer, which in turn necessitates a proper key management and storage for this encryption.

3.6 Example: Email address pseudonymization

An e-mail address constitutes a typical identifier of an individual. An e-mail address has the form local@domain, where the local part corresponds to the user that owns the address, and the domain corresponds to the mail service provider.

Users tend to use the same e-mail address for different applications, sharing it with various organizations, e.g. when they sign up for online accounts. Moreover, e-mail addresses are often published online. Due to these special characteristics, when e-mail addresses are used as identifiers, their protection is especially important.

In this use case, email addresses are considered as identifiers (e.g. in a database or online service), while analyzing the application of different pseudonymization techniques to them. It is always considered that the pseudonymization process is performed by a pseudonymization entity (e.g. data controller) as part of the operation/provision of a service.

Counter and random number generator

Both counter and RNG can be used for the pseudonymization of emails with the use of a mapping table. Pseudonymization is strong as long as the mapping table is secured and stored separately from the pseudonymized data.

Example of email address pseudonymization with RNG or counter (full pseudonymization)

	number generator	Pseudonym (counter generator)
lan@abc.eu	328	10
minh@wxyz.com	105	11
ngoc@abc.eu	209	12
duy@qed.edu	83	13
lan@wxyz.com	512	14
dao@clm.eu	289	15

In this example both counter and RNG result to pseudonyms that do not reveal any information on the initial identifiers (email addresses) and do not allow any further analysis (e.g. statistical analysis) on the pseudonyms. In order to increase utility, it is possible to apply pseudonymization only to a part of the email address, e.g. the local part.

Example of email address pseudonymization with RNG or counter (only local part pseudonymization)

Email address	Pseudonym (Random number generator)	Pseudonym (counter generator)
lan@abc.eu	328@abc.eu	10@abc.eu
minh@wxyz.com	105@wxyz.com	11@wxyz.com
ngoc@abc.eu	209@abc.eu	12@abc.eu
duy@qed.edu	83@qed.edu	13@qed.edu
lan@wxyz.com	512@wxyz.com	14@wxyz.com
dao@clm.eu	289@clm.eu	15@clm.eu

While the emails are pseudonymized, it is still possible to know the domain and, thus, conduct relevant analysis (e.g., number of email users originating from the same domain). As discussed, counter may be weaker in terms of protection as it allows for predictions due to its sequential nature (e.g. in cases where email addresses come from the same domain, the use of counter may reveal information regarding the sequence of the different email users in the database).

Starting from this case, depending on the level of data protection and utility that the pseudonymization entity needs to achieve, different variations might be possible by retaining different levels of information in the pseudonyms (e.g. on identical domains, local parts, etc.).

Examples of email address pseudonymization with RNG - various utility levels

Email address	Pseudonym (RNG) retaining the info on identical domains	Pseudonym (RNG) retaining also the info on identical country/extension	Pseudonym (RNG) retaining the info on identical local parts and domains	Pseudonym (RNG) retaining the info on identical country/extension, domains and local parts
lan@abc.eu	328@1051	328@1051.3	328@1051	328@1051.3
minh@wxyz.com	105@833	105@833.7	105@833	105@833.7
ngoc@abc.eu	209@1051	209@1051.3	209@1051	209@1051.3
duy@qed.edu	83@420	83@420.8	83@420	83@420.8
lan@wxyz.com	512@833	512@833.7	328@833	328@833.7
dao@clm.eu	289@2105	289@2105.3	289@2105	289@2105.3

The main pitfalls of both counter and RNG lie with the scalability of the technique in cases of large datasets, especially if it is required that the same pseudonym is always assigned to the same address (i.e., in a deterministic pseudonymization). In such case, the pseudonymization entity needs to perform a cross-check throughout the whole pseudonymization table whenever a new entry is to be pseudonymized. Complexity increases in more sophisticated cases of implementation as those shown above (e.g. when the pseudonymization entity needs to classify email addresses with the same domain or the same country without revealing this domain/country).

Cryptographic hash function

The total number of worldwide email accounts is roughly estimated to 4.7 billion ≈ 232 (since, despite the theoretically practically infinite size of the valid email addresses space, existing addresses lie in a much smaller space). This makes email addresses easily found or guessed, thus rendering cryptographic hash functions a weak technique for pseudonymization. Indeed, it is trivial to any insider or external adversary, having access to a pseudonymized list of email addresses, to perform a dictionary attack. This observation is relevant to all pseudonymization scenarios (independently of whether the pseudonymization entity is the controller, the processor or a trusted third party).

Reversing an e-mail address from its hash value

Despite the aforementioned pitfalls of cryptographic hash functions, it should be pointed out that, as indicated, service providers often share email addresses with third parties, just by simply hashing them. A concrete example of such case is the operation of the so-called custom audience lists, which provides to companies the possibility to compare hashed values of customers' email addresses for defining common lists of customers.

Notwithstanding the above significant data-protection risks, the cryptographic hash values could still be of some use under certain conditions, e.g., for internal coding of email addresses (such as for example in the context of research activities) and as validation/integrity mechanism for a data controller. Hash functions could also be used to pseudonymize parts of an email address (e.g., only the domain part), thus allowing some utility on the derived pseudonyms; if the remaining part is pseudonymized by a stronger method (e.g. MAC), then the risk of reversing the whole initial e-mail address is significantly reduced.

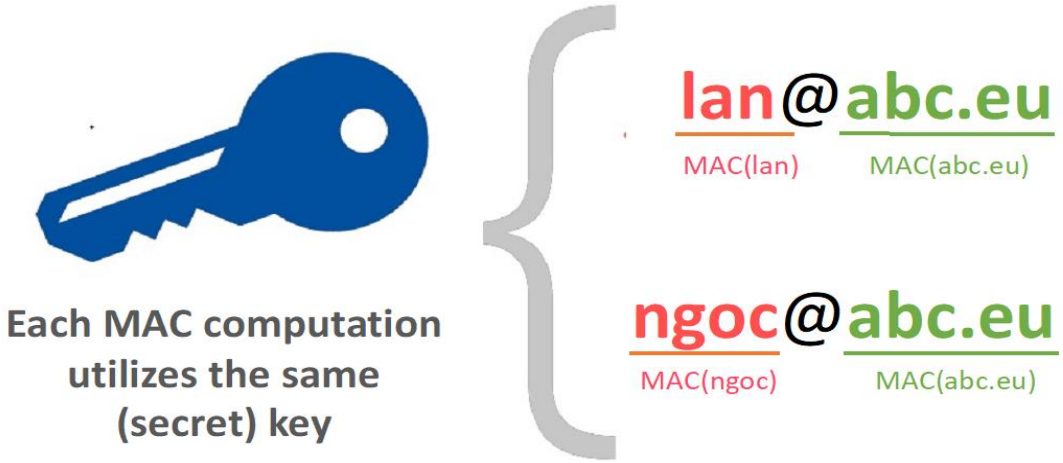
Message authentication code

Compared to simple hashing, a message authentication code (MAC) provides significant data protection advantages also for email address pseudonymization, as long as the secret key is securely stored. Moreover, the pseudonymization entity may use different secret keys, for different sectors, to generate for example different sector-based pseudonyms for the same e- mail address. A MAC can also be used to restrict the controller from having access to the email addresses in cases where access to the pseudonyms is sufficient for the particular purpose of processing. Such a case could be, for example,

in interest-based display advertising, in which the advertisers need to associate a unique pseudonym for each individual but without being able to reveal the user's original identity.

As in previous techniques, in order to increase utility of the pseudonyms, different implementation scenarios could be discussed in practice. For example, one possible approach would be to apply the MAC separately to different parts of the e-mail address (e.g., local and domain parts), using the same secret key. A characteristic example is shown below: the usage of the same key for each MAC results in generating the same sub-pseudonyms for the corresponding domain parts (in green color) whenever the email address domains are identical. However, since the output of a MAC has a fixed size, which is generally much larger than the size of the initial e-mail address, the resulting pseudonyms may be of quite large size (which is further increased if different parts are pseudonymized separately).

Using MAC to generate pseudonymized e-mail addresses with some utility



One important aspect regarding practical implementation of MAC is recovery. It should be stressed that even the data pseudonymization entity, which has access to the secret key, is not able to directly reverse the pseudonyms; such a reversion can be obtained only indirectly, by re- producing the pseudonyms for each known e-mail address in order to see the matches with the pseudonymized list. If a pseudonymization mapping table is available, reversing pseudonyms is trivial, but in such a case, the storage requirements also increase. For these reasons, MAC is probably not the most practical pseudonymization technique in cases that the data controller needs to be able to map pseudonyms to e-mail addresses easily.

Encryption

An alternative to MAC is encryption, applied especially in a deterministic way, i.e., by utilizing a secret key to produce a pseudonym for each e-mail address (symmetric encryption). Deployment is more practical in such case, since there is no need to provide for a pseudonymization mapping table: recovery is directly possible through the decryption process.

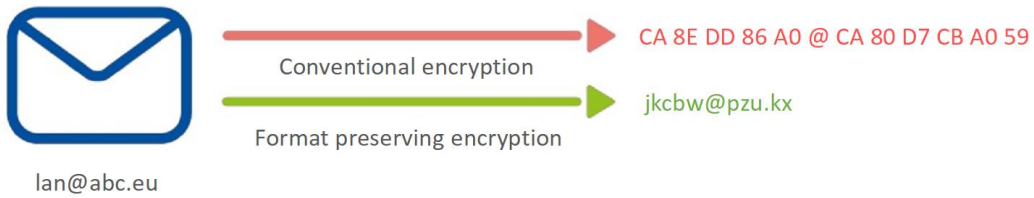
Note that, although some asymmetric (public key) cryptographic algorithms can be implemented in a deterministic way, they are not recommended for the pseudonymization of e-mail addresses (or for other data types). For example, assumed that the pseudonymization entity needs to generate, for each e-mail address, different pseudonyms for different – internal or external – users/recipients (with the assumption that each recipient will be able to re-identify his or her own data but not the pseudonymized data of other recipients). One possibility to achieve this goal would be to encrypt the emails with the public key of each recipient, thus allowing only the specific recipient to perform the decryption. Assuming that the public keys are in principle available to anyone, any adversary may mount a dictionary attack based on known (or guessed) e-mail addresses.

The nature of encryption by default does not allow for utility of the pseudonymized data. Encrypting separately the parts of an e-mail address may suffice to alleviate this issue, similarly to the message authentication codes, in which the MAC can be replaced by an encryption algorithm. Generally, to allow pseudonyms to carry some useful information, specific cryptographic techniques can be used; an illustrative example is given with format preserving encryption.

Format preserving encryption (FPE)

A database scheme might expect a particular data type for specific fields. For example, an e- mail address is expected to contain a local part (info), followed by an @ symbol, which in turn is followed by a domain. If there is no need, for the data controller, to retain the initial e-mail addresses but there is still need to keep a pseudonymized list by keeping the structure of the database, format preserving encryption is a way for achieving this. There are several known implementations on format-preserving encryption, based on known encryption schemes. In any case, any (pseudo)random substitution of characters by other characters lying in the same alphabet - i.e., the set of alphanumeric characters enriched by special characters appearing in local parts of e-mail addresses - suffices to ensure that the derived pseudonym has the desired form. The difference between FPE and conventional cryptography see below.

Conventional vs. format preserving encryption to derive pseudonym from e-mail address



Note that a symmetric stream cipher has been used for the conventional encryption, in order to ensure that the derived pseudonym has the same length with the initial address (the characters of the derived pseudonym are non-alphanumeric and, thus, are given in the hexadecimal form).

Depending on the case, it might be needed to appropriately engineer FPE implementations, in order to avoid the emergence of patterns that may leak information on the individuals' identities.

4 CONCLUSIONS

European Union General Data Protection Regulation, APPI, PDPA and/or other national Personal Data Protection Regulations and national laws requires to put in place appropriate technical and organizational measures to implement the data protection principles and safeguard individual rights. This is ‘data protection by design and by default’. It means integration or ‘bake in’ data protection into all processing activities and business practices, from the design stage right through the lifecycle

Data protection by design is about considering data protection and privacy issues upfront in everything we do. It helps to ensure that the compliance with GDPR’s fundamental principles and requirements, and forms part of the focus on accountability.

Key is to take organizational approach to ensure certain outcomes FPT Software must consider data protection issues as part of the design and implementation of systems, services, products and business practices

FPT Software must make data protection to an essential component of the core functionality of your processing systems, applications and services.

FPT Software shall process only personal data needed in relation to the purposes(s), and only use the data for those purposes

FPT Software shall protect personal data automatically in any IT system, service, product, and/or business practice, so that individuals should not have to take any specific action to protect their privacy.

By default, and by design FPT Software must take measures to protect personal data. In the design phase, blueprint phase data minimization, encryption, pseudonymization or masking must be considered. That means, wherever it is possible personal data must be pseudonymized, encrypted or masked.

All software architects, project manager und BU leads must consider data minimization, encryption, pseudonymization or masking (e.g., UI masking SAP) in every development project by default and design.

5 APPENDIX

5.1 Definitions

Abbreviations	Description
PII, Personal Identifiable Information, Personal Data	Refer to the personal data defined by the EU GDPR (Article 4 (1)), 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.
Data Subject	EU GDPR (Article 4 - 1), Data subject refers to any individual person who can be identified, directly or indirectly.
Data Controller	EU GDPR (Article 4 - 7), Data Controller means the natural or legal person, public authority, agency or anybody which alone or jointly with others, determines the purpose and means of processing of personal data; where the purpose and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law.
Data Processor	EU GDPR (Article 4 - 8), Data Processor means a natural or legal person, public authority, agency or anybody which processes data on behalf of the controller.
Recipient	EU GDPR (Article 4 - 9), A natural or legal person, public authority, agency or anybody, to which the personal data are disclosed, whether third party or not.
Third Party	EU GDPR (Article 4 - 10), A natural or legal person, public authority, agency or anybody other than the data subject, controller, processor and persons who under direct authority of controller or processor, are authorized to process personal data
Pseudonymization	Pseudonymization is the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person (GDPR, art. 4(5).
Anonymization	Anonymization is a process by which personal data is irreversibly altered in such a way that a data subject can no longer be identified directly or

Abbreviations	Description
	Indirectly, either by the data controller alone or in collaboration with any other party (ISO/TS 25237:2017).
Identifier	Identifier is a value that identifies an element within an identification scheme. A unique identifier is associated to only one element. It is assumed that unique identifiers are used, which are associated to personal data.
Pseudonym	Pseudonym, also known as cryptonym or just nym, is a piece of information associated to an identifier of an individual or any other kind of personal data (e.g., location data). Pseudonyms may have different degrees of link ability (to the original identifiers). The degree of link ability of different pseudonym types is important to consider for evaluating the strength of pseudonyms but also for the design of pseudonymous systems where a certain degree of link ability may be desired (e.g., when analyzing pseudonymous log files or for reputation systems).
Pseudonymization function	Pseudonymization function, denoted P, is a function that substitutes an identifier Id by a pseudonym $pseudo$.
Pseudonymization secret	Pseudonymization secret, denoted s is an (optional) parameter of a pseudonymization function P. The function P cannot be evaluated/computed if s is unknown.
Recovery function	Recovery function, denoted R, is a function that substitutes a pseudonym $pseudo$ by the identifier Id using the pseudonymization secret s. It inverts the pseudonymization function P.
Pseudonymization mapping table	Pseudonymization mapping table is a representation of the action of the pseudonymization function. It associates each identifier to its corresponding pseudonym. Depending on the pseudonymization function P, the pseudonymization mapping table may be the pseudonymization secret or part of it.
Pseudonymization entity	Pseudonymization entity is the entity responsible of processing identifiers into pseudonyms using the pseudonymization function. It can be a data controller, a data processor (performing pseudonymization on behalf of a controller), a trusted third party or a data subject, depending on the pseudonymization scenario. It should be stressed that, following this definition, the role of the pseudonymization entity is strictly relevant to the practical implementation of pseudonymization. The responsibility for the whole pseudonymization process (and for the whole data processing operation in general) always rests with the controller.
Identifier domain / pseudonym domain	Identifier domain / pseudonym domain refer to the domains from which the identifier and the pseudonym are drawn. They can be different or the same domains. They can be finite or infinite domains.

Abbreviations	Description
Adversary	Adversary is an entity that tries to break pseudonymization and link a pseudonym (or a pseudonymized dataset) back to the pseudonym holder.
Re-identification attack	Re-identification attack is an attack to pseudonymization performed by an adversary that aims to re-identify the holder of a pseudonym.
DPO/GDPO	Data Protection Officer/Global Data Protection Officer
DPIA	Data Protection Impacted Assessment
PIMS	Personal Information Management System
EU	European Union

5.2 Related Documents

No	Code	Name of documents
1	EU GDPR	EU General Data Protection Regulation
2	95/46/EC	EU Data Protection Directive 95/46/EC
3	Privacy shield	EU-U.S. and Swiss-U.S. Privacy Shield Frameworks designed by the U.S. Department of Commerce and the European Commission and Swiss Administration to provide companies on both sides of the Atlantic with a mechanism to comply with data protection requirements when transferring personal data from the European Union and Switzerland to the United States in support of transatlantic commerce.
4	APPI	Act on the Protection of Personal Information, Japan. It came into force on 30 May 2017.
5	PDPA	Personal Data Protection Act 2012, Singapore
6	PDPO	Personal Data (Privacy) Ordinance, Hongkong, 2012
7	PIPA	South Korea’s substantial Personal Information Protection Act (PIPA) was enacted on Sept. 30, 2011
8	PIPEDA	Personal Information Protection and Electronic Documents Act, Canada 2018
9	Privacy Act, APPs, CDR	Privacy act Australia including Australian Privacy Principles, Consumer Data Right
10	HITRUST	Health Information Trust Alliance (CSF, Common Security Framework)
11	HIPAA	Health Insurance Portability and Accountability Act of 1996 (HIPAA), US
12	PCI DSS	Payment Card Industry Data Security Standard, May 2018
13	CCPA	California Consumer Privacy Act of 2018, Cal. Civ. Code §§ 1798.100 et seq.
14	VCDPA	Virginia Consumer Data Protection Act, 01/2023
15	PDPL, UAE	Decree-Law No. 45 of 2021
16	DPA Philippines	Republic Act 10173, Data privacy Act 2012
17	PIPL	Personal Information Protection Law of the People’s Republic of China and related laws and regulations
18	PDPA Thailand	Thailand’s Personal Data Protection Act, 06/2022

No	Code	Name of documents
19	PDPA Malaysia	Personal Data Protection Act 2010, Malaysia
20	TISAX	Trusted information security assessment exchange
21	BS10012: 2017	British Standard Personal Information Management System
22		Vietnamese laws on Privacy: - Article 21 of the 2013 Constitution - Article 38 of the Civil Code 2015 - Article 125 of the Penal Code - Clause 2 of Article 19 of the Labor Code Decree of the Vietnamese Government: Nghị Định Quy Định Về Bảo Vệ Dữ Liệu Cá Nhân Came in force 07/2023
23	FPT Software Personal Data Protection Handbook	PDP_ Handbook_Version_V3.4

5.3 Data Protection Law, Vietnam, Overview

There is no single data protection law in Vietnam. Regulations on data protection and privacy can be found in various legal instruments. The right of privacy and right of reputation, dignity and honour and fundamental principles of such rights are currently provided for in Constitution 2013 ("**Constitution**") and Civil Code 2015 ("**Civil Code**") as inviolable and protected by law.

Regarding personal data , the guiding principles on collection, storage, use, process, disclosure or transfer of personal information are specified in the following main laws and documents:

- **Criminal Code** No. 100/2015/QH13, passed by the National Assembly on 27 November 2015
- Law No. 24/2018/QH14 on Cybersecurity, passed by the National Assembly on 12 June 2018 ("**Cybersecurity Law**");
- Law No. 86/2015/QH13 on Network Information Security, passed by the National Assembly on 19 November 2015; as amended by Law No. 35/2018/QH14 dated 20 November 2018, on amendments to some articles concerning planning of 37 Laws ("**Network Information Security Law**");
- Law No. 59/2010/QH12 on Protection of Consumers' Rights, passed by the National Assembly on 17 November 2010; as amended by Law No.35/2018/QH14 dated 20 November 2018, on amendments to some articles concerning planning of 37 Laws ("**CRPL**");
- Law No. 67/2006/QH11 on Information Technology, passed by the National Assembly on 29 June 2006; as amended by Law No. 21/2017/QH14 dated 14 November 2017 on planning ("**IT Law**");
- Law No. 51/2005/QH11 on E-transactions, passed by the National Assembly on 29 November 2005 ("**E-transactions Law**");
- Decree No. 85/2016/ND-CP dated 1 July 2016, on the security of information systems by classification ("**Decree 85**");
- Decree No. 72/2013/ND-CP dated 15 July 2013 of the Government, on management, provision and use of Internet services and online information; as amended by Decree No. 27/2018/ND-CP dated 1 March 2018 and Decree No.150/2018/ND-CP dated 7 November 2018 ("**Decree 72**");
- Decree No. 52/2013/ND-CP dated 16 May 2013 of the Government; as amended by Decree No. 08/2018/ND-CP dated 15 January 2018, on amendments to certain Decrees related to business conditions under state management of the Ministry of Industry and Trade and Decree No. 85/2021/ND-CP dated 25 September 2021 ("**Decree 52**");
- Decree No. 15/2020/ND-CP of the Government dated 3 February 2020 on penalties for administrative violations against regulations on postal services, telecommunications, radio frequencies, information technology and electronic transactions ("**Decree 15**");
- Circular No. 03/2017/TT-BTTTT of the Ministry of Information and Communications dated 24 April 2017 on guidelines for Decree 85 ("**Circular 03**");
- Circular No. 20/2017/TT-BTTTT dated 12 September 2017 of the Ministry of Information and Communications, providing for Regulations on coordinating and responding to information security incidents nationwide ("**Circular 20**");
- Circular No. 38/2016/TT-BTTTT dated 26 December 2016 of the Ministry of Information and Communications, detailing cross-border provision of public information ("**Circular 38**");

- Circular No. 24/2015/TT-BTTTT dated 18 August 2015 of the Ministry of Information and Communications, providing for the management and use of Internet resources, as amended by Circular No. 06/2019/TT-BTTTT dated 19 July 2019 ("**Circular 25**"); and
- Decision No. 05/2017/QD-TTg of the Prime Minister dated 16 March 2017 on emergency response plans to ensure national cyber-information security ("**Decision 05**").

Applicability of the legal documents will depend on the factual context of each case, e.g businesses in the banking and finance, education, healthcare sectors may be subject to specialized data protection regulations, not to mention to regulations on employees' personal information as provided in Labour Code 2019 ("**Labour Code**").

The most important Vietnamese legal documents regulating data protection are the Cybersecurity Law and Network Information Security Law. Cybersecurity laws in other jurisdictions that were inspired by the GDPR of the EU, the Cybersecurity Law of Vietnam shares similarities with China's Cybersecurity Law enacted in 2017. The law focuses on providing the government with the ability to control the flow of information. The Network Information Security Law enforces data privacy rights for individual data subjects.

A draft Decree detailing a number of articles of the Cybersecurity Law ("**Draft Cybersecurity Decree**"), notably including implementation guidelines for data localization requirements, together with a draft Decree detailing the order of and procedures for application of a number of cybersecurity assurance measures and a draft Decision of the Prime Minister promulgating a List of information systems important for national security, are being prepared by the Ministry of Public Security ("**MPS**") in coordination with other relevant ministries, ministerial-level agencies and bodies.

MPS has drafted a Decree on personal data protection ("**Draft PDPD**"), which is contemplated to consolidate all data protection laws and regulations into one comprehensive data protection law as well as make significant additions and improvements to the existing regulations. The Draft PDPD was released for public comments in February 2021 and was originally scheduled to take effect by December 2021. The Finalization process consuming much more time than the MPS first anticipated. PDPD was finalized and was coming in force 07/2023.