A Methodology for Assuring Privacy by Design in Information Systems

Siham Arfaoui , Abdellatif Mezrioui and Abdelhamid BelmekkI

National Institute of Posts and Telecommunications, Morocco

Abstract: There is no doubt that privacy by design PbD has become a structuring paradigm for personal data protection. Certainly this paradigm has been in use since 1995; however the GDRP "The General Data Protection Regulation", by considering PbD in 2018 as a legal obligation, it testifies the PbD seven principles relevance. Companies are therefore called to put in place technical and organizational measures to integrate PbD into companies. Hence the need for a methodology to provide an exhaustive approach adapted to this implementation. Given the focus of the literature on the implementation of methodologies dedicated to the embodiment of PbD only in software systems, this article aims to propose an ISPM methodology "Information System Privacy Methodology" which focuses on the implementation of PbD in the enterprises architecture, specifically in information systems taking into account all the technical and organizational aspects which must be adopted for the said goal success.

Keywords: Privacy, Togaf, Information Systems, ISPM, Linddun, Threat Modeling, Privacy Enhancing Technologies PET, Personal Data.

1. Introduction

In the era where digital has found its place in every aspect of our daily lives, the needs and habits evolution become linked to digital innovations. Information technologies IT and business have therefore become two close partners. Companies have found support in their information system IS for efficient work organization. The search for consistency, control and collective action visibility are ensured by technologies integration in the information system IS. The digital transformation has thus made it possible to increase the company performance by developing the customer and the employee experience. If no one denies the prowess that digitalization promises for the IS, the debate on privacy digitization threat has been growing in recent years. A privacy invasion in an IS thus implies unauthorized access to personal information or unauthorized collection, use or communication of such information. Some of privacy breaches occur when personal information is stolen, lost or disclosed in error. A privacy breach can also result from a procedural error or an operational failure.

In 2018, the GDRP law "The General Data Protection Regulation" was published, instituting the individuals' protection with regard to the personal data processing. Article 25 of the GDRP entitled "Data protection by design and by default" incorporates the paradigm of "Privacy by design". The principle of data protection by design means that the company must integrate personal data protection from projects design related to the company's data processing. This concept stems from a 1995 report on the Privacy-Enhancing Technologies PET of a joint team made up of the Information and Privacy Commissioner of Ontario (Canada), Ann Cavoukian, of the Protection Authority Data from the Netherlands and the Netherlands Organization for Applied Scientific Research.

Hence the deal with the privacy protection issue should not only consider a separate software system but rather the entire IS. Thus, integrating PbD into enterprise architecture in a comprehensive way requires adopting a coherent approach. The chosen approach must be adapted to the PbD implementation specifications and must also take into consideration all IS layers: process, data flow, applications and technical architecture. The methodologies available in the literature [1, 2, 3, 4, 5…] are dedicated to PbD implementation only in software systems. Software systems are just one of the building blocks of an IS. Hence this article interest: it proposes a methodology called ISPM "Information System Privacy Methodology" which aims PbD implementation in IS. In the following section, the authors have addressed the requirements that ISPM must meet. They then synthesized in the related works section the methodologies best known in the literature dealing with the implementation of PbD. Thus, the presentation of the ISPM methodology in a synthesized way is done before detailing it step by step in the last section.

2. Background: Privacy by design

PbD or privacy from the design stage is a system engineering approach that considers privacy throughout the process [6]. The close alignment between the "Security by Design" and "Privacy by Design" work was introduced in January 2013 [7]. Privacy by design is based on seven fundamental principles [8]:

1. Proactive not reactive: it seeks to anticipate and prevent privacy-invasive events before they happen by not waiting for privacy risks to materialize.
2. Privacy as the default setting: it seeks to build privacy measures directly into any given information, communication technology system and business practice by default.
3. Privacy embedded into design: it seeks to embed privacy into IS design and the architecture and business practices. It does not bolt it on after the fact.
4. Positive-Sum: it seeks to accommodate all legitimate interests and objectives in a positive-sum "win-win" way, not through a zero-sum approach involving unnecessary trade-offs.

5. End-to-End security: it seeks to ensure cradle-to-grave secure information lifecycle management, end-to-end.
6. Visibility and transparency: it seeks to keep IS component parts and business practices operations visible and transparent to users.
7. Respect for user privacy: it seeks to respect and protect interests of the individual, above all. It keeps it user centric.

3. Requirements

Many of ISs do not comply with the GDRP requirements. Then, bringing an IS from a state that does not comply with the GDRP to a compliant state needs an adequate methodological approach. The objectives of this methodology would be:

- Identify and census of all personal data that is manipulated in the IS in question.
- Identify the actors who will be able to manipulate personal data and grant them the appropriate rights to do so.
- Prevent other actors from manipulating personal data by establishing control rules and using adequate anonymization techniques.
- Also, IS evolves and therefore the methodological approach must take into account this fact of scalability and be invoked continuously.

The methodology should assume that the ISs could be in one of the following states:

- They are already designed and deployed and do not comply with the GDRP, therefore they must be adapted to comply with this regulation.
- They are to be designed from scratch respecting the GDRP.
- They are already designed and comply with the GDRP but they must evolve while respecting the GDRP.

To summarize, the methodology to be designed must be:

- PbD driven methodology, which implies taking into account all PbD principles.
- Holistic, so it must take into account the process layer, data flow layer, the application layer and the technical architecture.
- Integrated with the IS design method used in the company and / or proposes generic and standardized models.
- In addition, it should carry out a privacy treats analysis, their impact and occurrence degrees.
- Ultimately, it should propose strategies for resolving threats or at least for reducing them considerably.

4. Related Works

In 2008, Seiya Miyazaki et al. have published the PRET methodology [1, 2] which instead of being based on threats identification; it is based on a tool that allows listing a number of personal data protection recommendations by domain and by country. PRET is based on the laws and the legislations in force. Seiya et al. recognize that PRET covers a limited laws number. In addition, PRET does not offer technical mechanisms and solutions that can be implemented to put the collected recommendations into practice.

In the same year 2008, the PRIS methodology was published [3]; it covers security and privacy protection implementation. PRIS is objective oriented, more precisely; it provides a set of concept for modeling confidentiality requirements allowing the requirements translation into a system model. PRIS is however a methodology with a considerable abstraction level and which does not study privacy threats in any detailed and explicit way.

In May 2010, Fahriya Seda Güres published her thesis [4] on a methodology dedicated to personal data protection called MPRA. The strong point of this methodology is that it is oriented towards personal data collection. MPRA also gives great importance to stakeholder’s analysis [4] which must be involved in personal data protection. Thus, transgressions analysis in terms of personal data use is recommended to decide which PETs to use. However, this methodology lacks in identifying and modeling personal data threats.

OASIS, acronym for the Organization for Advancement of Structured Information Standards, which is a global consortium, published the PMRM methodology in 2012 [9]. It provides an approach for developing operational solutions related to confidentiality issues. PMRM focuses on the importance of stakeholder analysis, specifies the categories recommendations to be considered. After an association establishment, the mapping between the processes and the technical mechanisms is carried out to develop architecture and implement it. This methodology comes, for the first time, to take into account the "personal data flow" aspect. However, there is personal data threats identification and modeling lack.

In 2014, ProPan Problem-Based Privacy Analysis was published [5]. This approach aims to identify privacy threats when analyzing software system requirements. ProPan does not rely entirely on privacy analyst but allows computer-assisted threat identification derived from the relationships between the stakeholders, the technology and the system's personal information. Propan is based on analysis threats recommendations in an automated way, it also remains among the pioneering methods to integrate threats analysis related to privacy and to model them in threat trees. However, it should be noted that ProPan does not cover the passage from threat analysis to the choice of the PET to be used in the implementation phase.

In 2015, Kim Wuyts published her thesis on a methodology for implementing privacy protection in software systems called Linddun [10]. Linddun is an acronym for the threats it studies, which are: Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of information, Unawareness, Non compliance. Linddun adheres to the PbD paradigm and aims to help software engineers with limited privacy expertise to introduce it into the life cycle of software creation [11]. Linddun is based on a threat analysis, which follows software system modeling using data flow diagram DFD. Linddun models threats is an interesting practical tree [12] that is imbued with Microsoft's STRIDE methodology [13, 14]. Linddun also offers a transition from threat analysis to choosing the solution to adopt [15] while prioritizing identified threats [11].Linddun is therefore a methodology that covers a significant number of the methodological privacy implementation aspects applied to software systems. In addition, Kim continues to refine Linddun by works [16] published in July 2018. In 2019, other works were published for the same purpose by her [17].

However, Linddun can be criticized for the choice of working with the Data Flow Diagram DFD as an entry point for the analysis process and the system understanding. In fact, in her thesis [11], Kim did not benchmark modeling standards to decide the computer system modeling diagram choice. She just justified her DFD use by the fact that it is the model used by Security Development Lifecycle SDL [13]. SDL is the method she based on to develop Linddun. It should be mentioned that since the release of the mature version of SDL in 2004, other modeling standards have been created like Business Process Model and Notation BPMN which was adopted by the OMG in 2006. Indeed, after the publication of SDL and well before Linddun publication, BPMN was made available to the public in 2008. BPMN allows a better understanding of the modeled system. It provides easy-to-use notation that is independent of the implementation environment. In addition, BPMN has been adopted since 2013 as an international standard ISO / IEC 19510.In her publication [16], Kim admits that the elemental analysis offered by the DFD diagram need to be improved by a context analysis based on input, process and output. All the same she continues to use the DFD diagram. Another thing to reproach Linddun for is that Kim just hinted at the need to start by understanding the system to be modeled without providing recommendations to help Linddun user in this mission.

As seen, the methodologies available in the literature [1, 2, 3, 4, and 5] are dedicated to PbD implementation only in software systems. Software systems are just one of the building blocks of an IS. This article goal is to propose a methodology for implementing privacy by design in IS, which can only be achieved by modeling the IS in question. It was not until the 80s to see the emergence of the first methods and concepts of enterprise architecture modeling. IBM, at the time leader of the IT market, promote the Business Systems Planning BSP method considered as a pioneer of enterprise architecture. John Zachmman, who is one of the leading designers of IBM's BSP method and employee, had used the term "Enterprise Architecture" for the first time in 1982, before presenting the first version of Information Systems Architecture Framework in 1987, which is named more commonly as Zachmman method. Over time, the Zachmman framework has inspired other methods [18]. The authors will not detail all the approaches of enterprise architecture modeling and their historical evolutions, nor will be able to make a complete listing, because at least 80 architectural frameworks have been recorded around the world [18]. Some count only one user, their designer only, while others have several tens of thousands users. For example, The Open Group, which publishes the TOGAF framework, indicates in real time the number of people certified [18].

Togaf success is due to its need satisfaction to have a common framework facilitating architectural practices capitalization, more precisely, Togaf is positioned as a generic method which brings together a set of techniques centered on the transformation of enterprise architecture [19], this approach is called Architecture Development Method ADM, it integrates into its process the different facets: strategy, business, technique, governance and change planning. How to go from the initial architecture to the target architecture? The answer to this question is at the heart of the Togaf framework. It offers a set of diagrams and matrices making it possible to target the changes to be adopted and to document them [19].

TOGAF standard reflects an architecture capability structure and content within an enterprise [20]. Capability is an ability that an organization, person, or system possesses. Capabilities are typically expressed in general and high-level terms and typically require a combination of organization, people, processes, and technology to achieve like marketing, customer contact, or outbound telemarketing.

5. Information System Privacy Methodology ISPM: overview

In what follows, the authors present the ISPM methodology approach. It consists of six steps as shown in Figure 1:

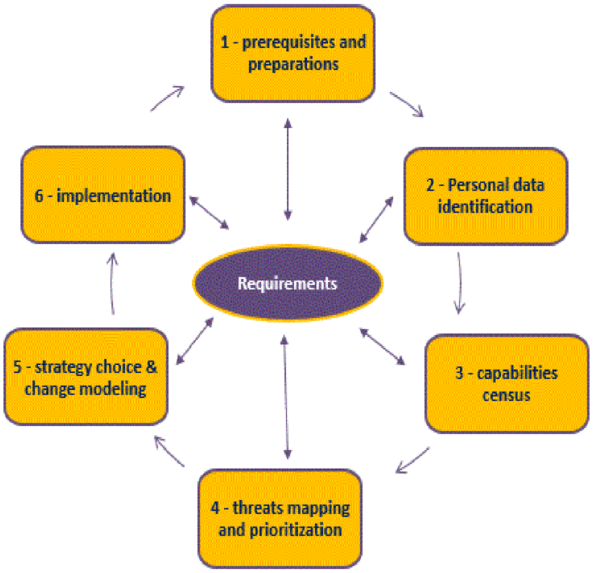
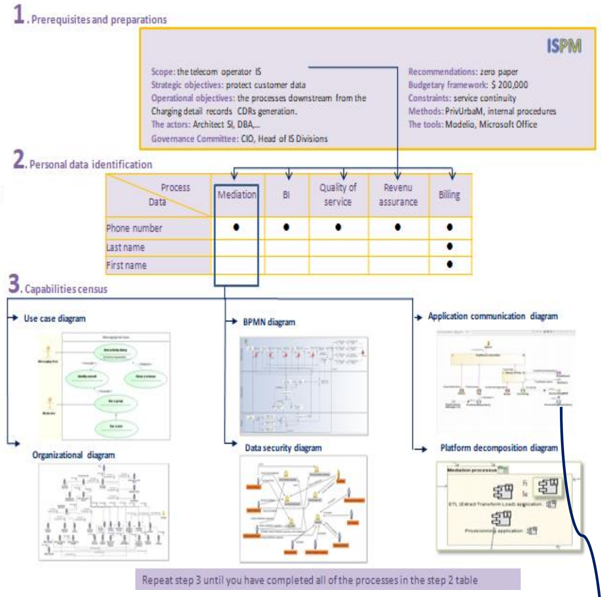

1 - prerequisites and preparations
2 - Personal data identification
3 - capabilities census
4 - threats mapping and prioritization
5 - strategy choice & change modeling
6 - implementation
Requirements

Figure 1. ISPM methodology

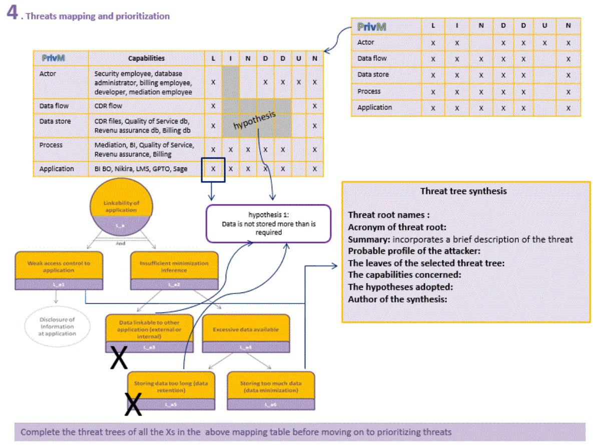
The first step objective is to determine the expectations from ISPM implementation. It also serves to define the context and the perimeter of the action field. Thus, the stakeholders are identified, also the constraints are synthesized and the recommendations are established. In step 2, an inventory of all personal data, sensitive personal data and quasi-identifiers is carried out concerning the chosen perimeter or in the entire information system. The input of this step is data classification according to personal data protection laws. As for its output, it represents all of personal data and processes of the chosen scope. The purpose of step 3, for its part, is to identify all the chosen perimeter capabilities. These are the different processes, the data flows, the application layer, the technical architecture and related stakeholders that interact with personal data. The goal is to have an understanding of what exists to know where to act. The input of this third step is existing knowledge based on the paperwork, employees, norms and standards of the company activity field. As for output, it is a set of standardized diagrams that have been carefully chosen by ISPM from the set of diagrams proposed by Togaf. Figure 2 summarizes the steps 1, 2 and 3. The fourth step aims to rule on all threats of the study scope. These threats prioritization is subsequently developed. The input to this step is the trees proposed by Linddun and completed by the authors adding threats tree. As for output, it represents all the threats to be prioritized and for which the appropriate strategy must be determined to remedy them. Figure 3 summarizes the steps 4. It is at step five that the strategy to be followed is decided and also the whole change to operate on the different processes, the data flows, the application layer, the technical architecture and the rights of the stakeholders are ruled. Step 5 is summarized in figure 4 and 5. The choice of appropriate privacy enhancing technologies to implement the changes agreed in the different bricks of the perimeter constitutes the sixth and last step of ISPM approach using the matrix made available for this purpose.

In what follows, we will illustrate the different steps of ISPM in figures 2, 3, 4 and 5.


1. Prerequisites and preparations
ISPM
Scope: the telecom operator IS
Strategic objectives: protect customer data
Operational objectives: the processes downstream from the Charging detail records CDRs generation.
The actors: Architect SI, DBA,...
Governance Committee: CIO, Head of IS Divisions
Recommendations: zero paper
Budgetary framework: $ 200,000
Constraints: service continuity
Methods: PrivUrbaM, internal procedures
The tools: Modelio, Microsoft Office
2. Personal data identification
3. Capabilities census
Use case diagram
BPMN diagram
Application communication diagram
Organizational diagram
Data security diagram
Platform decomposition diagram
Repeat step 3 until you have completed all of the processes in the step 2 table

Process / Data	Mediation	BI	Quality of service	Revenu assurance	Billing
Phone number	●	●	●	●	●
Last name					●
First name					●

Figure 2. The steps 1, 2 and 3 of the ISPM methodology


4. Threats mapping and prioritization
hypothesis
Linkability of application
L_a
And
Weak access control to application
L_a1
Insufficient minimization inference
L_a2
Disclosure of information at application
Data linkable to other application (external or internal)
L_a3
Excessive data available
L_a4
Storing data too long (data retention)
L_a5
Storing too much data (data minimization)
L_a6
X
X
hypothesis 1:
Data is not stored more than is required
Complete the threat trees of all the Xs in the above mapping table before moving on to prioritizing threats

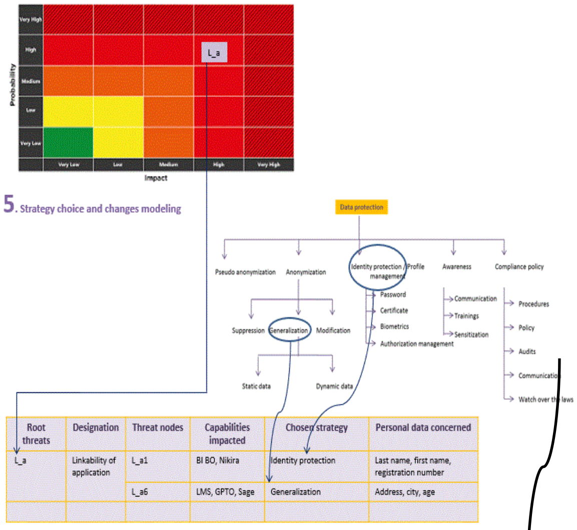
PrivM	Capabilities	L	I	N	D	D	U	N
Actor	Security employee, database administrator, billing employee, developer, mediation employee	x			x	x	x	x
Data flow	CDR flow	x						x
Data store	CDR files, Quality of Service db, Revenu assurance db, Billing db	x						x
Process	Mediation, BI, Quality of Service, Revenu assurance, Billing	x	x	x	x	x		x
Application	BI BO, Nikira, LMS, GPTO, Sage	x	x	x	x	x		x

PrivM	L	I	N	D	D	U	N
Actor	x	x		x	x	x	x
Data flow	x	x	x	x	x		x
Data store	x	x	x	x	x		x
Process	x	x	x	x	x		x
Application	x	x	x	x	x		x

Threat tree synthesis

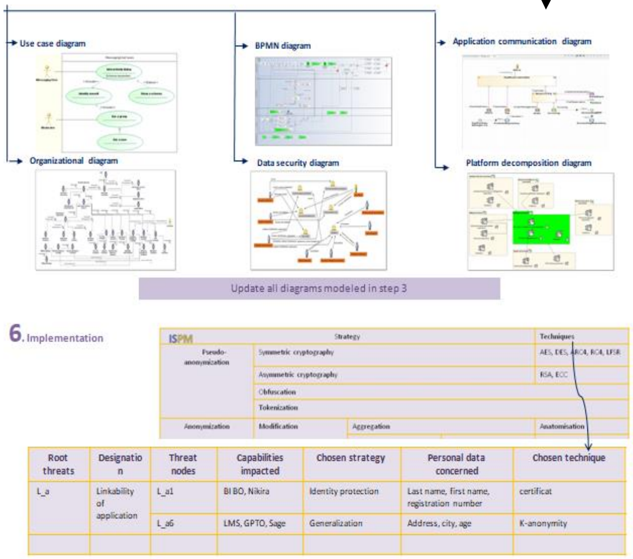
Threat root names :
Acronym of threat root:
Summary: incorporates a brief description of the threat
Probable profile of the attacker:
The leaves of the selected threat tree:
The capabilities concerned:
The hypotheses adopted:
Author of the synthesis:

Figure 3. The steps 4 of the ISPM methodology



Root threats	Designation	Threat nodes	Capabilities impacted	Chosen strategy	Personal data concerned
L_a	Linkability of application	L_a1	BI BO, Nikira	Identity protection	Last name, first name, registration number
		L_a6	LMS, GPTO, Sage	Generalization	Address, city, age

Figure 4. The steps 5 of the ISPM methodology


Use case diagram
BPMN diagram
Application communication diagram
Organizational diagram
Data security diagram
Platform decomposition diagram
Update all diagrams modeled in step 3
6. Implementation

ISPM	Strategy		Techniques
Pseudo-anonymization	Symmetric cryptography		AES, DES, ARC4, RC4, LFSR
	Asymmetric cryptography		RSA, ECC
	Obfuscation		
	Tokenization		
Anonymization	Modification	Aggregation	Anatomisation

Root threats	Designation	Threat nodes	Capabilities impacted	Chosen strategy	Personal data concerned	Chosen technique
L_a	Linkability of application	L_a1	BI BO, Nikira	Identity protection	Last name, first name, registration number	certificat
		L_a6	LMS, GPTO, Sage	Generalization	Address, city, age	K-anonymity

Figure 5: The step 5 continuity and the step 6 of the ISPM methodology

6. Information System Privacy Methodology ISPM: step by step

6.1. Step 1: prerequisites and preparations

The purpose of this phase is to prepare the company wishing to protect its personal data to know the crosscutting activities related to this project governance. This stage is triggered by the decision to start work. The bricks that need to be considered are:

- The scope: depends on the company in question choice. It can be either the entire IS, as it can be defined as a set of processes, data flows, personal data, applications or technical architecture.
- Objectives: they are often known before starting work. In deciding to initiate the personal data protection, the strategic objectives must be broken down into milestones over time.

  The responsibilities allocation is a key step in finalizing objectives definition, as it implies negotiation on the objectives feasibility, and obtaining an agreement between the stakeholders.
- The actors: this involves identifying the participants in the personal data protection project, their influence on commitments, their essential concerns that must be addressed and taken into account during this project development. The participant's role and their tasks should also be determined.
- It is also necessary to designate a governance committee which will be responsible for the following points:
  - Create and manage the project.
  - Control and validate the solutions implemented.
  - Guarantee the consistency and convergence of the solutions chosen with the company's strategy.
  - Manage conflicts.
  - Ensure regular activities monitoring and report to management.
- Recommendations: they depend on the company in question, however two recommendations are intrinsic to any project should be taken into consideration for to implementing ISPM methodology:
  - The need for the company in question to go paperless. Dematerialization or zero paper is the replacement paper documents by computer files and digital data management (exchange, storage, archiving ...). In addition of generating significant savings benefit, reducing the time associated with administrative procedures and therefore optimizing the responsiveness and efficiency of the business, zero paper implementation also guarantees data integrity and reliability, facilitates its archiving and saving, keeps the flow traceability but above all ensure exchanges confidentiality.
  - The need for documentation of all the elements mentioned in this section and of all business activities. The purpose of such approach is essentially to ensure the accuracy and reliability of the information to be used on the project, and thus to have the guarantee that all operations are carried out in accordance with what is planned and wrote.
- The budgetary framework: depends on the company concerned and its Capital Expenditure CAPEX management strategy.
- Constraints: in whatever project, there are constraints. Hiding them face is a flaw that must be excluded as a governance committee. This notion of constraint requires a precise, rigorous analysis upstream of the project. The basic constraints are budget, deadlines, resources and Ensuring service continuity. Depending on these constraints and priorities, the most appropriate combination is proposed.
- Methods: in our case, understanding the ISPM methodology and applying it while using internal procedures.
- Tools: To carry out their missions, stakeholders must use a wide range of tools. The right tool for the right use ... The choice will depend on those used by the company for usual projects management. If there is no tools for setting objectives, actions organization, planning, resource management, diagrams modeling, we recommend that a benchmark be carried out before deciding which one to adopt.

Note that communication plays a central role during this phase; it is a question of ensuring the common stakeholders understanding in order to obtain a consensus of the directions and expected results. At the end of this phase, everything must be validated in order to start the project:

- The perimeter
- Orientation: a consensus of principles, objectives, major requirements as well as constraints.
- The organization: the actors, their roles, their respective implications.
- The roadmap.
- A macroscopic vision on the architecture of the IS.

In short, at the end of this stage, the need is to know where to go, how and who are going.

6.2. Step 2: Personal data identification

At this stage, the personal data, the sensitive personal data and the Quasi-identifiers of the study chosen perimeter in step 1 must be identified. Remember that personal data is all data that can allow direct or indirect person identification. Personal data can also be classified as sensitive data such as medical records, biometric records such as DNA, fingerprints, iris, retina, voice, etc. For quasi-identifiers, they are non-explicit identifiers such as date of birth, gender, postal code, etc. These

quasi-identifiers, if combined, can help infer the identity of the person in question.
In all of the following, we will use "personal data" to include personal data, sensitive personal data and Quasi-identifiers.
To help the interested party to identify all personal data, the authors propose below a categorization of personal data based on personal data protection laws [21].The actors must try to find all personal data referring to the different categories and classification of personal data mentioned below relating to the chosen scope. Meetings should be scheduled with employees of the chosen perimeter to conduct a study to determine all personal data and special attention should be paid to the exhaustive inventory of Quasi-identifiers specific to the field in question. For example, in the telecom domain, International Mobile Subscriber Identity IMSI or International Mobile Equipment Identity IMEI are Quasi-identifiers that only a person in the telecom domain can know about their existence. For this reason that it is necessary to be equipped during the census personal data meetings with the categorization below and to instruct the employees of the perimeter in question to be exhaustive by citing all data which help directly and indirectly to identify a person (in the example cited, when mentioning the category "identification of the persons concerned", the employee of the perimeter in question must cited IMSI and IMEI).

The categories of personal data are:

- Identification of the persons concerned: surname, first name, address, date and place of birth, email, telephone number, national identity number, photos, etc.
- Behavior: Consumption habits, geographic location, lifestyle, leisure
- Professional data: CV, training, diplomas, experiences…
- Financial situation: bank information, income, debts, bank card number, amount in the bank account, loans made
- Sensitive data: racial or ethnic origin, political opinions, religious or philosophical beliefs, union membership, health data, genetic data, data relating to security measures, biometric data, etc.

Once all personal data are identified following the meetings carried out with the employees in question in an iterative way, it is necessary to draw up the matrix of table 1 that represents the exhaustiveness of the chosen perimeter personal data correlated to the different processes where they appear.

Table 1. *Example of a personal data census matrix*

Process Data	Process 1	Process 2	Process 3	Process 4	Process 5
Personal data 1	●	●	●	●	●
Personal data 2					●
Personal data 3					●

Once the processes are delimited exhaustively and definitively using the census of personal data, step 3 of the ISPM methodology consists on identifying the capabilities of the processes resulting from step 2. The authors will detail this section in the followed section.

6.3. Step 3: capabilities census

In step 3, the goal is to facilitate the IS understanding and allow the exhaustive inventory of the processes, data flows, applications and components of the technical architecture on which it is necessary to act to set up privacy by design. This objective is ensured by the ability to model the collected processes of step 2 based on Togaf. The term "Capabilities" comes from Togaf, it means the capacity that an organization, a person or a system has: "capabilities" are generally expressed in general and high level terms and generally require a combination of organization, people, process and technology [20]. Togaf has a panoply of diagrams and matrices called artifact allowing modeling capabilities. The authors have meticulously chosen six diagrams to ensure a modeling assuring to define the necessary different perspectives; the chosen artifacts include processes, data flow, applications and technical architecture modeling.
In what follows, the authors will approach the six diagrams carefully chosen from the panoply of diagrams proposed by Togaf [22]:

- Use case diagram: this is a diagram chosen to give a global vision of the studied business process functional behavior, which will help to understand the system. A use case represents a discrete interaction unit between a user (human or machine) and a system. It is a significant unit of work. In a use case diagram, users are called actors, they interact with use cases.
- Organizational diagram: it gives an idea of the different actors, their hierarchical relationships and their interactions. It is essential to determine the actors that should be integrated in stage 4 of ISPM.
- BPMN business process diagram: this diagram details the different stages by which a process is executed. It shows the "who", the "what", the "when", the "where" and the "how" and helps to analyze the "why". It is essential for mastering the business changes to be made to integrate privacy by design into the process in question.
- Data security diagram: Data constitutes an essential element of the company's heritage. Ensuring data security means that its integrity will not be compromised, and that their access authorization will be properly controlled. This diagram helps to delimit different actors' perimeter interacting with data and makes it possible to determine each actor attributions grant.
- Application communication diagram: The information system is essentially based on a set of communicating computer applications. The gradual increase in the number and diversity of these applications and their interrelationships within companies has made their control more and more complex. In order to manage their development, IT managers must rely on application maps. The application communication diagram gives visibility on all the applications of the IS or of the process in question and their interactions.

- Platform decomposition diagram: it represents the technological platform that supports the operations of Information Systems Architecture.

These six diagrams are modeled in a gradual and iterative way for all of the processes identified in step 2. These diagrams will be completed during the modeling of all the processes of step 2. At the end of step 3, at least six diagrams will constitute the output of this section. The same diagram can represent two or more processes from step 2; the nesting of the same diagram for different processes is possible when necessary. Note that a diagram can be refined and completed by the set of processes, which gives a synthetic and global vision of the diagram in question (often the case of the data security diagram).

Thus, all the capabilities of the chosen perimeter: people, processes, applications and technologies are all cited in the set of the discussed modeled diagrams. The next section shows how to use these capabilities census.

6.4. Step 4: threats mapping and prioritization

ISPM considers the same threat families as Linddun. Linddun stepped into the identification of these threats families from the work of Pfitzmann and Hansen [23] for the categories Linkability, identifiability and detectability, while "disclosure of information" is steeped in the STRIDE methodology [13]. Non-repudiation, Unawareness and Non-compliance come from the experience of Kim [14] and her team in the field of personal data protection.

The adapted ISPM methodology threat mapping imbued with Linddun is represented in the table 2 which includes an additional layer: the application layer added to the other capabilities that Linddun identifies.

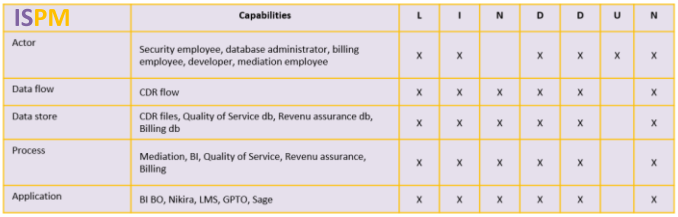
Table 2. Threats Correlated to Capabilities - ISPM

	L	I	N	D	D	U	N
Actor	X	X		X	X	X	X
Data flow	X	X	X	X	X		X
Data store	X	X	X	X	X		X
Process	X	X	X	X	X		X
Application	X	X	X	X	X		X

Note that regarding the "Actor" capability named "user" in Linddun, the authors added the consideration of "Detectability, Disclosure of information and non-compliance" risk not integrated by Linddun. The consideration of these three threats in the “Actor” category is justified by the fact that an actor can potentially distinguish whether information of interest is present or not, can also expose information to unauthorized third parties to see it. He also may not be in line with the requirements and procedures put in place by the company with regard to the protection of personal data. The only threat not integrated for the case of "Actor" is "Non repudiation" because the actor can in all circumstances deny having done an action, this denial can only be contradicted if we have recourse to the verification of the data flow, data store or the application in question.

It must be then add the mapping of the table 2 with the capabilities identified in step 3 to create a custom mapping table that corresponds to the system analyzed. The resulting mapping table must contain all the capabilities generated from the modeling performed in step 3. This table must then be used as a checklist throughout the analysis, because each "X" in the table represents a potential threat to a specific capability generated from the modeling performed in step 3. Therefore, each "X" must be documented as a threat, if not applicable, a hypothesis must be explicitly written to explain the reasons of not maintaining the associated threat and in this case the cell is grayed out. Table 3 illustrates this mapping:

Table 3. Threat mapping with capabilities



ISPM	Capabilities	L	I	N	D	D	U	N
Actor	Security employee, database administrator, billing employee, developer, mediation employee	X	X		X	X	X	X
Data flow	CDR flow	X	X	X	X	X		X
Data store	CDR files, Quality of Service db, Revenu assurance db, Billing db	X	X	X	X	X		X
Process	Mediation, BI, Quality of Service, Revenu assurance, Billing	X	X	X	X	X		X
Application	BI BO, Nikira, LMS, GPTO, Sage	X	X	X	X	X		X

The case where hypotheses rule out threats is illustrated in table 4 (gray cells)

Table 4. Example of threat mapping with capabilities taking into account hypotheses

ISPM	Capabilities	L	I	N	D	D	U	N
Actor	Security employee, database administrator, billing employee, developer, mediation employee	X	▨		X	X	X	X
Data flow	CDR flow	X	▨	▨	▨	▨		X
Data store	CDR files, Quality of Service db, Revenu assurance db, Billing db	X	▨	▨	▨	▨		X
Process	Mediation, BI, Quality of Service, Revenu assurance, Billing	X	X	X	X	X		X
Application	BI BO, Nikira, LMS, GPTO, Sage	X	X	X	X	X		X

The cells marked in gray in the figure above are potential threats that were considered not relevant to the specific use scenario of the system studied. Each cell indicated by an “X” shows that there will be a threat to privacy at the level of the corresponding element of the information system. These are the threats that we will actually consider. Note that when you decide to reject the "X's", you should always document this decision as an explicit assumption. In theory, each "X" should be examined (and documented) individually. In practice, however, it is advisable to apply the "reduction" technique. This implies that more than one "X" can be combined when applied to the same threat. This is possible for the "X" that involve elements of the same type of information system (for example, data flow or process) and when the threat that corresponds to the "X" is the same, because it involves the same type of data (e.g. usernames and passwords, whether or not etc. By combining these "Xs”, the resulting threat description document will become easier to manage. However, an exception to the rule stipulates that for "X" of the same type of "non-compliance" element can be combined. Indeed, threats of "non-compliance" are rather generic and apply to the whole system. All the “Xs” of “non-compliance” can be combined and treated together (this decision must also be documented as an assumption).

For each cell marked in step 4 mapping (see table 2), a threat tree exists. The privacy threat trees are inspired from the Security Development Lifecycle SDL [13] and based on the most recent confidentiality watches. These threat trees reflect common attack patterns and help people think about privacy conditions in the information system. The threat trees being regularly updated, the authors refer the reader to the latest version of the trees, available on the Linddun website [24].Note that we have added our own trees on the non-existing application layer in Linddun (6 trees, figures 9, 10, 11, 12, 13 and 14) and also 2 non-existing trees in the “Actor” layer for the “detectability and disclosure of information” threats. Regarding the non-compliance of actor threat tree, the one produced by Kim [24] incorporates the non-compliance of actor, embodied in “attacker tampering with privacy policies and makes consents inconsistent” [12] even if in her works does not include this threat. Note that "Entity" and "Actor" represent the same capability. In what follows, the authors will present the 9 trees added. They start by presenting the generic legend of the added trees. Figure 6 summarizes all the symbols used.

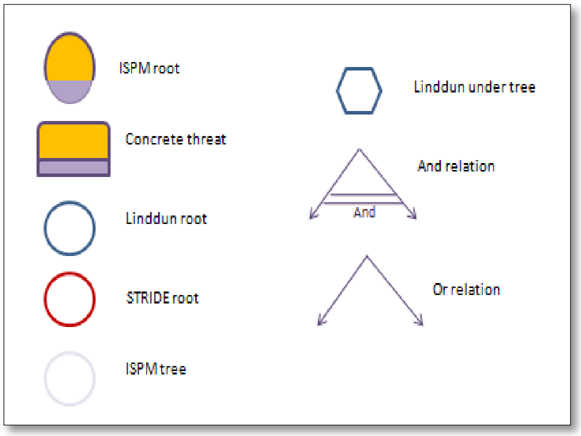


Figure 6. Legend of added threat trees

Beginning with the detectability and disclosure of information from actor trees (figure 7 and 8):

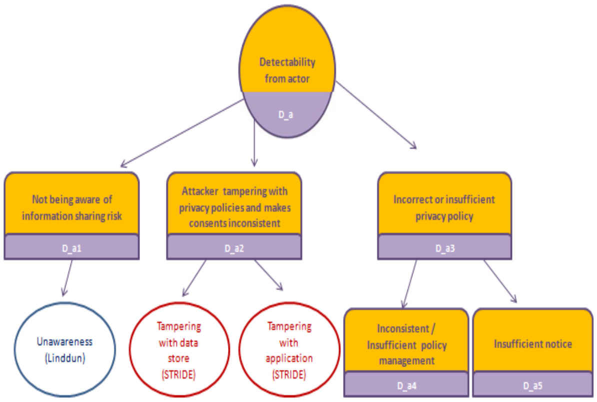

Detectability from actor
D_a
Not being aware of information sharing risk
D_a1
Attacker tampering with privacy policies and makes consents inconsistent
D_a2
Incorrect or insufficient privacy policy
D_a3
Unawareness (Linddun)
Tampering with data store (STRIDE)
Tampering with application (STRIDE)
Inconsistent / Insufficient policy management
D_a4
Insufficient notice
D_a5

Figure 7. Threat tree "detectability from actor"

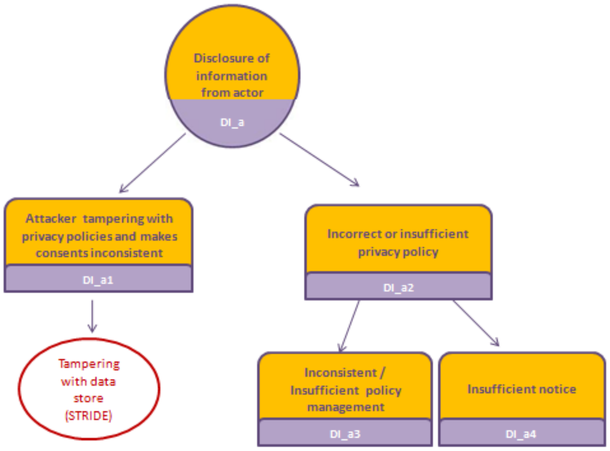

Disclosure of information from actor
DI_a
Attacker tampering with privacy policies and makes consents inconsistent
DI_a1
Incorrect or insufficient privacy policy
DI_a2
Tampering with data store (STRIDE)
Inconsistent / Insufficient policy management
DI_a3
Insufficient notice
DI_a4

Figure 8. "Disclosure of information from actor" threat tree

Now passing to the application layer threat trees for the six threat families as agreed in table 2.

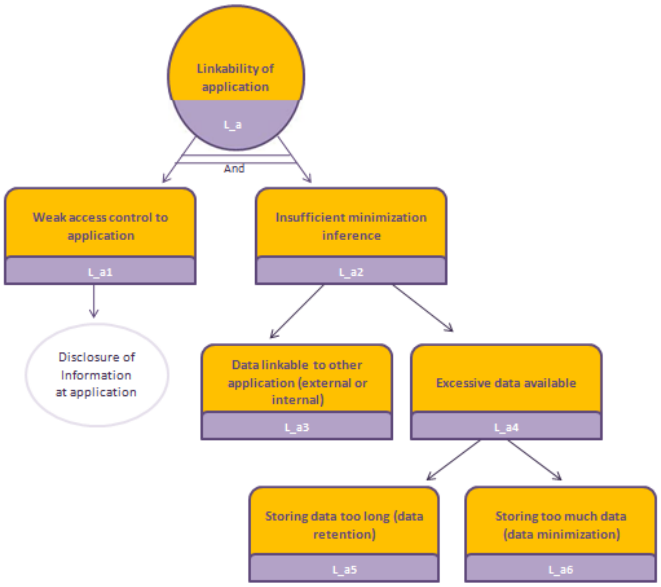

Linkability of application
L_a
And
Weak access control to application
L_a1
Insufficient minimization inference
L_a2
Disclosure of Information at application
Data linkable to other application (external or internal)
L_a3
Excessive data available
L_a4
Storing data too long (data retention)
L_a5
Storing too much data (data minimization)
L_a6

Figure 9. "Linkability of application" threat tree

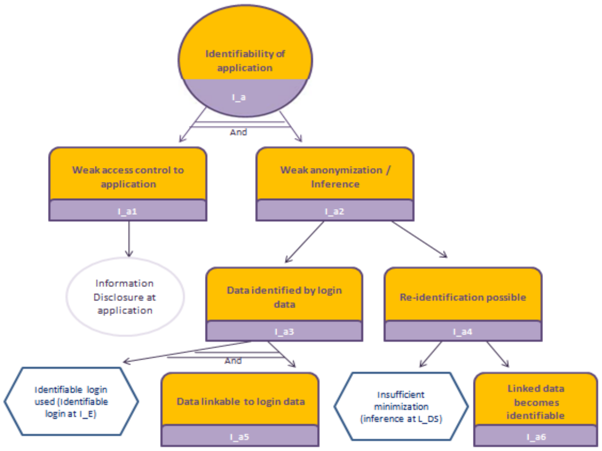

Identifiability of application
I_a
And
Weak access control to application
I_a1
Weak anonymization / Inference
I_a2
Information Disclosure at application
Data identified by login data
I_a3
Re-identification possible
I_a4
And
Identifiable login used (Identifiable login at I_E)
Data linkable to login data
I_a5
Insufficient minimization (inference at L_DS)
Linked data becomes identifiable
I_a6

Figure 10. "Identifiability of application" threat tree

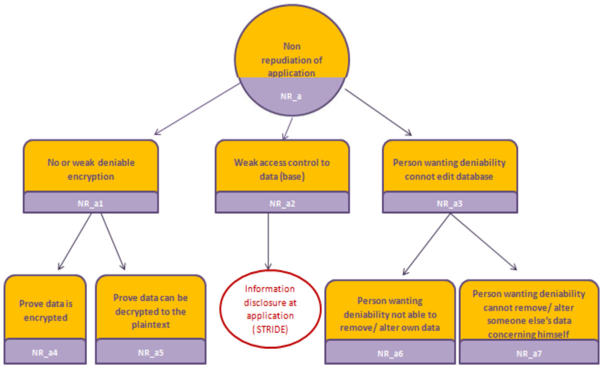

Non repudiation of application
NR_a
No or weak deniable encryption
NR_a1
Weak access control to data (base)
NR_a2
Person wanting deniability connot edit database
NR_a3
Prove data is encrypted
NR_a4
Prove data can be decrypted to the plaintext
NR_a5
Information disclosure at application (STRIDE)
Person wanting deniability not able to remove/ alter own data
NR_a6
Person wanting deniability cannot remove/ alter someone else's data concerning himself
NR_a7

Figure 11. "Non repudiation of application" threat tree

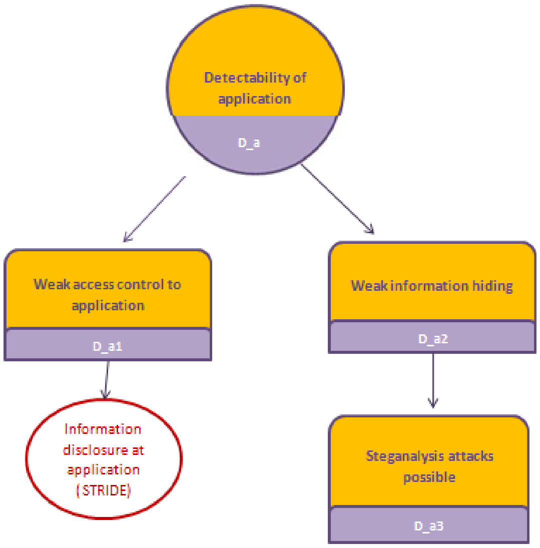

Detectability of application
D_a
Weak access control to application
D_a1
Weak information hiding
D_a2
Information disclosure at application (STRIDE)
Steganalysis attacks possible
D_a3

Figure 12. "Detectability of application" threat tree

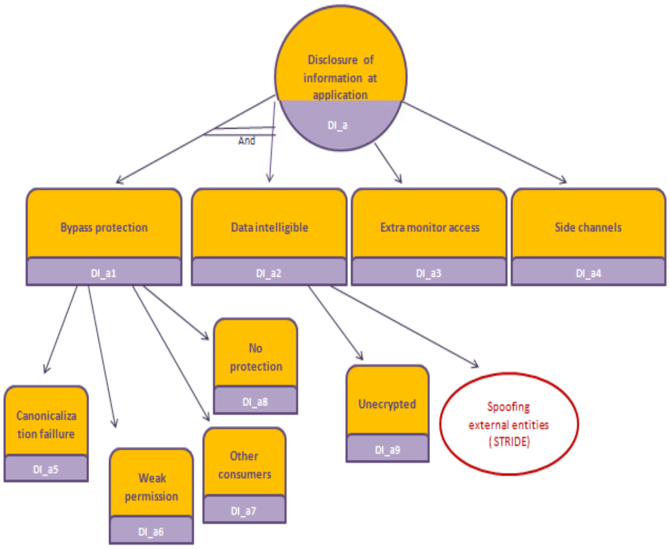


Figure 13. "Disclosure of information at application" threat tree

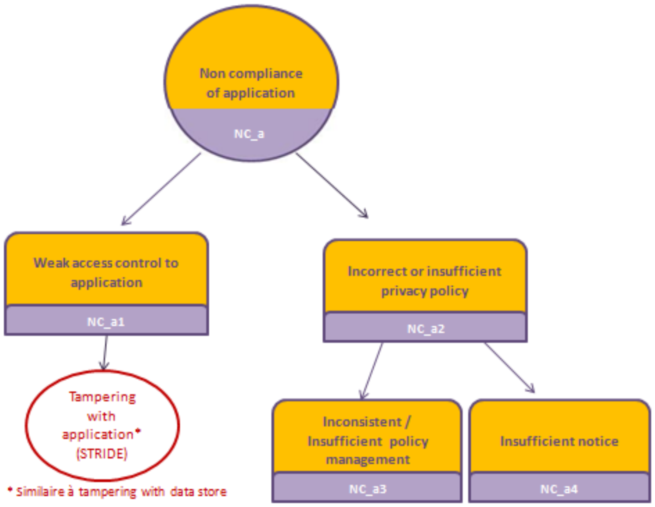
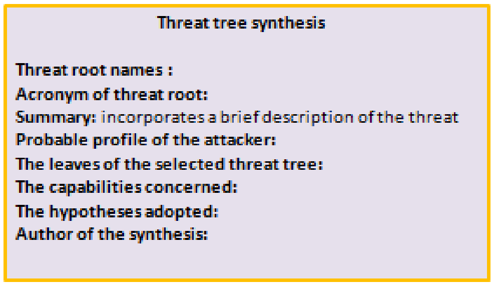

Non compliance of application
NC_a
Weak access control to application
NC_a1
Incorrect or insufficient privacy policy
NC_a2
Tampering with application* (STRIDE)
* Similaire à tampering with data store
Inconsistent / Insufficient policy management
NC_a3
Insufficient notice
NC_a4

Figure 14. Non compliance of application" threat tree

The result of this threat analysis phase is a set of threat scenarios that must be documented. For this purpose and for each threat tree root, the authors suggest filling the template shown in Figure 15:


Threat tree synthesis

Threat root names :
Acronym of threat root:
Summary: incorporates a brief description of the threat
Probable profile of the attacker:
The leaves of the selected threat tree:
The capabilities concerned:
The hypotheses adopted:
Author of the synthesis:

Figure 15. Threat tree synthesis

Before going ahead and looking for appropriate mitigation measures for identified threats, threats must be prioritized. Due to time or budget constraints, only the most significant threats will often be taken into account for inclusion in the specification of requirements and, therefore, in the personal data protection implementation solution. Risk assessment techniques support this step. In general, risk is calculated based on the probability of the attack scenario and its impact. The risk value is used to sort the threat roots: the higher the risk, the greater the threats root: Risk = probability x impact.

ISPM is independent of the risk assessment technique used. The analyst is free to choose the technique of his choice, for example the risk assessment methodology of the OWASP [25], DREAD of Microsoft [26], the special publication of NIST 800-30 [27] or OCTAVE from SEI [28]. These techniques exploit the information contained in the threat files, such as the capabilities involved (for the impact), the profile of the attacker as well (for the probability).

In what follows, the choice of the strategy allowing facing the threats should be selected.

6.5. Step 5: strategy choice and changes modeling

Arriving at this stage, it must be decided the strategy to adopt to mitigate the threats collected. In Figure 16, a summary of all the strategies to be adopted is presented.

More information about anonymization and pseudo-anonymization techniques could be found in [29, 30]

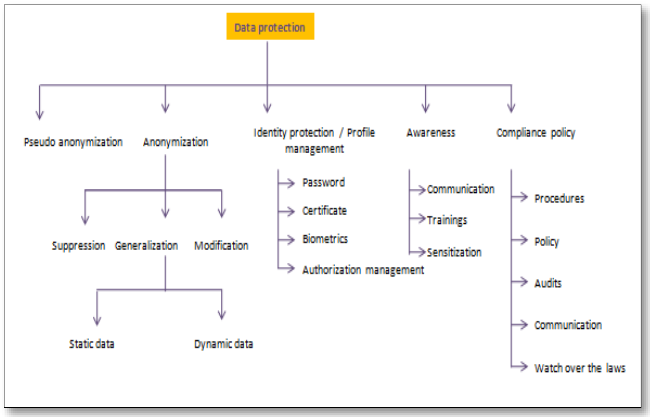


Figure 16. Personal data protection strategies

Once the strategy has been chosen, a summary table can be drawn up to facilitate monitoring of the decisions taken as shown in table 5:

Table 5. Summary of the chosen strategies

Root threats	Designation	Threat nodes	Capabilities impacted	Chosen strategy	Personal data concerned
L_a	Linkability of application	L_a1	BI BO, Nikira	Identity protection	Last name, first name, registration number
		L_a6	LMS, GPTO, Sage	Generalization	Address, city, age

The work carried out so far in step 5 cannot give rise to a complete solution for integrating privacy by design in an IS for three reasons:

- Even if the personal data protection strategy is choosed, it could take place to anonymize a data at a stage of the process and want to make the data clear for the rest of the process. This type of change can never be taken into account unless the updating of the modeling of the BPMN and use case diagrams already modeled in step 3,
- Choosing a strategy may involve adding a component to the technical architecture or an application that does not appear if the update to the modeling in step 3 is not made. Therefore updating the modeling done in step 3 is essential and will allow understanding and identifying all the necessary changes on all layers of the IS.
- The roles attribution and grant associated to each actor might be changed to guarantee the personal data protection integration. It will not appear if the security diagram is not updated.

The order of updating the diagrams is shown in Figure 17:

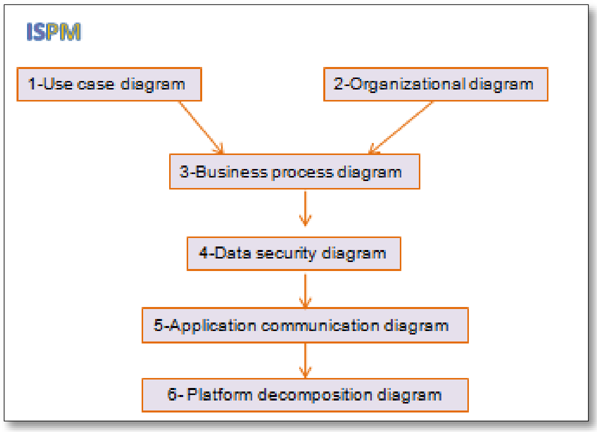

ISPM
1-Use case diagram
2-Organizational diagram
3-Business process diagram
4-Data security diagram
5-Application communication diagram
6- Platform decomposition diagram

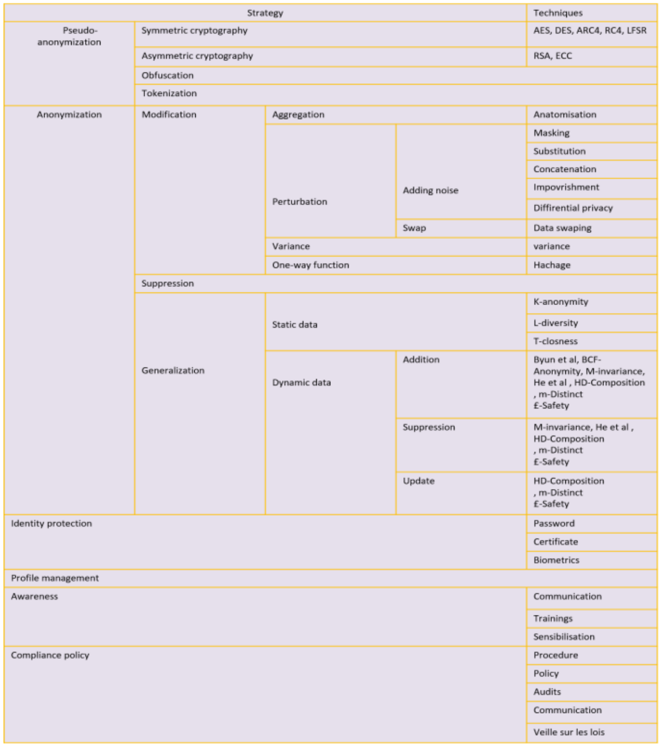
Figure 17. Diagram update order

Changes made on the old modeling can be added with another color on the chosen modeling tool to highlight them. At this stage, it remains to choose the techniques for implementing the decided strategies. The authors deal with this aspect in the next section.

6.6. Step 6: implementation

In order to be able to choose the implementation techniques from the panoply offered by the literature, the table 6 is drawn up and summarizes the techniques to be used in the different strategies that have been mentioned in Figure 16.

Table 6. Summary of personal data protection techniques



Strategy				Techniques
Pseudo-anonymization	Symmetric cryptography			AES, DES, ARC4, RC4, LFSR
	Asymmetric cryptography			RSA, ECC
	Obfuscation			
	Tokenization			
Anonymization	Modification	Aggregation		Anatomisation
		Perturbation	Adding noise	Masking
				Substitution
				Concatenation
				Impovrishment
				Diffirential privacy
			Swap	Data swaping
		Variance		variance
		One-way function		Hachage
	Suppression			
	Generalization	Static data		K-anonymity
				L-diversity
				T-closness
		Dynamic data	Addition	Byun et al, BCF-Anonymity, M-invariance, He et al , HD-Composition , m-Distinct £-Safety
			Suppression	M-invariance, He et al , HD-Composition , m-Distinct £-Safety
			Update	HD-Composition , m-Distinct £-Safety
Identity protection				Password
				Certificate
				Biometrics
Profile management				
Awareness				Communication
				Trainings
				Sensibilisation
Compliance policy				Procedure
				Policy
				Audits
				Communication
				Veille sur les lois

In addition, the authors have done a qualitative study of the anonymization techniques cited in table 6 [31] to help the interested party in choosing the appropriate technique to adopt. The study is based on four criteria: Identifiability, Linkability, inference and data quality.

7. Conclusion

The importance of integrating personal data protection into enterprise architecture is no longer questionable. Thus the information system of a company, which represents the bottleneck of data manipulation, needs to comply with the principles of privacy by design. To do this, an exhaustive methodological approach must be followed in order to ensure the efficiency of the IS adaptation to privacy by design in a lasting and evolutionary way throughout the company lifecycle. Hence the interest of this article that proposes a methodology for this objective called ISPM: Information System Privacy Methodology. ISPM was detailed step by step in this article and it fully meets the requirements that were set at the start of this article.

References

[1] Miyazaki, S., Mead,N., Zhan,J. Computer-Aided Privacy Requirements Elicitation Technique, published in IEEE Asia-Pacific Services Computing Conference, 2008. DOI 10.1109/APSCC.2008.263.

[2] Mead, N.R., Miyazaki,S., Zhan,J. Integrating Privacy Requirements Considerations into a Security Requirements Engineering Method and Tool published in Information Privacy, Security and Integrity journal, Vol. 1, No. 1, 2011.

[3] Kalloniatis,C., Kavakli,E., Gritzalis.S. Addressing privacy requirements in system design: The PriS method, DOI: 10.1007/s00766-008-0067-3 · Source: DBLP, 2008.

[4] Gürses,F.S. Multilateral Privacy Requirements Analysis in Online Social Network Services. Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering in the Arenberg School, 2010.

[5] Beckers,K., Faßbender,S., Heisel,M., Meis,R. A Problem-based Approach for Computer Aided Privacy Threat Identification. In Privacy Technologies and Policy, volume 8319 of LNCS, pages 1–16. Springer, 2014.

[6] Prosch, M. The Roadmap for Privacy by Design in Mobile Communications: A Practical Tool for Developers, Service Providers and Users, http://www.ontla.on.ca/library/repository/mon/25003/308516.pdf, December 2010.

[7] Cavoukian, A. Privacy and Security by Design: a convergence of Paradigms, published jointly by a team composed of the Information and Privacy Commissioner of Ontario, Ann Cavoukian and Oracle Corporation, http://www.discoveringidentity.com/2013/03/21/privacy-and-security-by-design-a-convergence-of-paradigms/, Mars 2013.

[8] Cavoukian, A., Dixon, M. Privacy and Security by Design: An Enterprise Architecture Approach, http://www.discoveringidentity.com/2013/09/23/privacy-and-security-by-design-an-enterprise-architecture-approach/, September 2013.

[9] Committee Specification, Privacy Management Reference Model and Methodology (PMRM) Version 1.0, 26, 2012.

[10] Wuyts,K. Privacy Threats in Software Architectures. Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering, 2015.

[11] Wuyts,K., Joosen,W. LINDDUN Privacy Threat Modeling: a tutorial, Technical Report (CW Reports), volume CW685, Department of Computer Science, KU Leuven, 2015.

[12] Wuyts,K., Scandariato,R., Joosen,W. LIND(D)UN privacy threat tree catalog, Report CW675, https://distrinet.cs.kuleuven.be/software/linddun. 2014.

[13] Howard,M., Lipner,S. The Security Development Lifecycle: A Process for Developing Demonstrably More Secure Software. Washington, USA: published by Microsoft Press a Division of Microsoft Corporation, 2006.

[14] Wuyts,K. LINDDUN: a privacy threat analysis Framework https://distrinet.cs.kuleuven.be/software/linddun, (2010).

[15] Deng,M., Wuyts,K., Scandariato,R., Preneel,B., Joosen,W. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements, 2010. DOI 10.1007/s00766-010-0115-7.

[16] Sion,L., Wuyts,K., Yskout,K., Landuyt,D.V., Joosen,W. Interaction-based Privacy Threat Elicitation, Published in 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), 2018. DOI: 10.1109/EuroSPW.2018.00017. 23-27.

[17] Sion,L., Landuyt,D.V., Wuyts,K., Joosen,W. Privacy Risk Assessment for Data Subject-aware Threat Modeling, Published in: 2019 IEEE Security and Privacy Workshops (SPW),2019.DOI: 10.1109/SPW.2019.00023.

[18] Pepin, J. Enterprise Architecture: mapping alignment business and applications of the information system, April 2018.

[19] The Open Group Standard. The TOGAF Standard, Version 9.2, 2018.

[20] Harrison, R,Togaf 9 Foundation Study Guide, Preparation for the Togaf Part 1 Examination,third edition, Oxford Brookes University, 2013.

[21] Declaration of Treatment Law N ° 09-08 promulgated by Dahir N ° 1-09-15 of 22 Safar 1430, February 18, 2009 - Article 15, https://www.cndp.ma/images/lois/CNDP-Declaration-Normale.pdf.

[22] The Open Group. Togaf version 9.1 Entreprise Edition :Sample Catalogs, Matrices and Diagrams v3, December 2011. Download from http://www.opengroup.org/bookstore/catalog/i093.htm.

[23] Pfitzmann,A. Hansen,M. A terminology for talking about privacy by data minimization: anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management (Version 0.33), technical report, TU Dresden and ULD Kiel, 2010. http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.

[24] Wuyts,K., Joosen,W. https://www.linddun.org/linddun-threat-catalog

[25] Ramadlan,M.F. OWASP, "Risk rating methodology", https://owasp.org/www-pdf-archive/Riskratingmanagement-170615172835.pdf. 2013.

[26] Chilton,S., Lee,M.J., Kiraly,F., Metcalf,H., Pang,W. Dread Risks, Journal of risk and uncertainty, volume 33, pages 165-182, 2006.

[27] Stoneburner,G., Goguen,A. Feringa,A. "Risk management guide for information technology systems, special publication 800-30," , 2002. Available: http://csrc.nist.gov/publications/nistpubs/800-30/sp800-30.pdf

[28] Krikken,R. The OCTAVE Risk Assessment Methodologies, Gartner Research ID: G00203984, 2010.

[29] Arfaoui, S., Belmekki, A., Mezrioui.A. Privacy increase on telecommunication processes. Published in: 2018 International Conference on Advanced Communication Technologies and Networking (CommNet). 2018. DOI: 10.1109/COMMNET.2018.8360266.

[30] El Ouazzani.Z, El Bakkali,H. A Classification of non-Cryptographic Anonymization Techniques ensuring Privacy in Big Data in International Journal of Communication Networks and Information Security (IJCNIS) Vol. 12, No. 1, April 2020

[31] Arfaoui, S., Belmekki, A., Mezrioui.A A Qualitative-Driven Study of Irreversible Data Anonymizing Techniques in Databases.In Proceedings of SITA conference (SITA'20). ACM, Rabat, Morocco, 2020. https://doi.org/10.1145/3419604.3419788.