A New Generic Representation for Modeling Privacy

Myriam Clouet, Thibaud Antignac, Mathilde Arnaud, Gabriel Pedroza, Julien Signoles
Université Paris-Saclay, CEA, List
Palaiseau, France
firstname.lastname@cea.fr

***Abstract*—Verifying how personal data is used in industrial applications is of utmost importance for evaluating privacy. Existing works propose many definitions of privacy and offer a wide variety of approaches for verifying privacy properties. Yet, as far as we know, there is currently no generic setting allowing to model all these concepts, as well as to express and verify related privacy properties, whilst taking into account the various abstraction levels involved when designing and implementing a computer system.**

This work proposes a generic representation for modeling privacy relying on a new classification scheme and a new ontology. These aim to be generic enough for modeling key privacy concepts and their relationships, as well as helping express and verify related properties at different levels of abstractions. Our representation of privacy can also be specialized into various existing ones, according to different criteria, whilst allowing us to instantiate it for mapping its generic elements to existing examples or concrete use cases.

***Index Terms*—privacy, classification, ontology, consent, modeling language**

1. Introduction

Using personal data is privacy-critical in many domains, such as web applications, voice recognition and authentication, or health care systems. It is thus interesting to have a generic framework to study privacy applicable to all those domains.

More and more regulations about personal data protection are emerging around the world, e.g. in Europe through GDPR [2] but also in Australia [3] and Japan [4], which are widely recognized as contributing to the protection of the privacy of individuals. Even if each has its own specific features — privacy is not viewed the same way in the United States, Europe or China — some ideas are common, especially for privacy in computer systems. For instance, legal and ethical principles require to verify that personal data are used correctly and only for the purpose to which the concerned party has consented.

However, developing a privacy-aware computer system is not straightforward, and demonstrating that it has no privacy flaws is even harder. Even large companies, such as Google, have already been fined for incorrect use of personal data [22]: the information provided for the granting of consent was not sufficiently clear and unambiguous, so the consent was judged invalid. Indeed, privacy is a vast and complex notion, so each stakeholder usually chooses to represent privacy according to its needs. For example, legal instances [2] often represent privacy as a set of principles, while tool providers for privacy protection may represent it as a set of properties that a computer system must enforce [12]. Unfortunately, this variety of representations makes comparisons between different approaches complicated. Furthermore, in a digital setting, some issues arise from the need of handling privacy all throughout the software life-cycle. For instance, GDPR advocates the principle of data protection by design and by default [2], which has a positive impact on individual's privacy. In particular, article 25 states that the data controller must apply appropriate measures throughout the system lifecycle, from the definition of the means for processing personal data, to the processing itself. However, privacy can hardly be represented and verified in the same way throughout the whole life-cycle. How is it possible to ensure consent at both design and implementation time, while being sure that this term is used with the same intent each time?

This paper proposes a new generic representation for modeling privacy. It aims at helping compare solutions for privacy at different stages of the system life-cycle. While being generic enough for taking into account very diverse privacy contexts, our new representation is yet specializable and instanstiable on concrete examples and use cases. More precisely, our contributions are:

- GePyR, a new privacy representation, as a classification scheme relying on privacy categories, that is generic and specializable;
- PyCO, an instantiable ontology that models key privacy elements and their relationship;
- examples of specialization and instantiation on examples from the literature.

GePyRand PyCOmay be use together, like in our running example, but also separately. For example, using just GePyRto focus on a classification and comparison of papers, and using PyCOto identify elements needed to take in account when verifying privacy-related properties.

Our models use our own graphical representations. Indeed, among the existing modeling languages of which we are aware, none is able to represent all of our concepts in a visual, clear and succinct way. For example, UML class diagrams do not allow us to represent in a clear way the different entities and their relations, while UML use cases or data flow diagrams are useful to represent most relations but cannot model composition ones, for example

an aggregation relation.

The paper is organized as follows: Section 2 introduces the running example used along this paper. Section 3 presents an overview of existing representations of privacy, as well as approaches enhancing privacy protection. Section 4 introduces GePyR, our new generic representation of privacy. Section 5 presents an example of specialization of its privacy classification scheme. Section 6 introduces PyCO, our new ontology of privacy context. Section 7 illustrates our key notions on existing examples and a possible methodology for using our contributions on the running example. Section 8 compares our work with other privacy ontologies. Section 9 concludes and discusses future works.

2. Running Example

This paper will use a running example at several places. The global picture of this running example considers an R&D engineer team in a French organization that wants to develop a new website relying on some user personal data for two purposes: website administration on one hand and marketing on another hand. Among others, the website uses the users's e-mail addresses to inform them when their subscription expires, but also to send them targeted advertising. We will use this running example in Sections 4 and 6. How our contributions can be used on this running example will be illustrated in Section 7.3.

3. Existing Representations

This section provides a short survey of some existing privacy representations and solutions that ensure various privacy properties.

Actors involved in the field of privacy actually propose many definitions and discussions with related concepts, such as personal data protection, of privacy, depending on their points of view, in various types of documents such as legal documents [2], [6], taxonomies [7]–[9], frameworks [10], [11] and terminologies [12]. This variety of sources leads to a diversity of representations. Additionally, representations may be heterogeneous, even in the very same type of documents [9]. For identifying our privacy representations, we only focus on papers [2], [6]–[12]. Indeed, all the papers that we have looked at and that provide solutions for ensuring privacy, always rely on one (or, sometimes, several) of these representations, even though not always explicitly.

Since we reuse definitions from existing papers, the introduced representations are sometimes quite close or even redundant. Here, we only categorize them. Section 4 introduce how our classification scheme solves this issue, while Section 5 show how it can be specialized to recover these existing representations.

The above-mentionned papers introduce seven different representations of privacy, which are (using the names given in the original papers): **principles**, **properties**, **harmful activities**, **threats**, **goals**, **vulnerabilities**, or **dimensions**.

These seven representations may be split into two main categories. On the one hand, there are representations viewing privacy as concepts to be preserved:

- **Principles:** privacy is viewed as a set of principles to be followed by actors using personal data [2], [6], [10]; e.g. the *lawfulness, fairness and transparency* principle corresponds to "Personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject" [2].
- **Properties:** privacy can be defined through a set of properties to preserve; e.g. the *Undetectability* property corresponds to "the attacker cannot sufficiently distinguish whether an item of interest exists or not" [12].
- **Goals:** privacy is seen as a set of goals to be fulfilled [9]; e.g. the *Choice/Consent* goal corresponds to "consumers are given the option to decide what personal information collected about them is to be used and whether it may be used for secondary purposes".

On the other hand, there are representations relying on concepts that threaten privacy:

- **Harmful Activities:** Solove identifies a set of harmful activities that can affect privacy; e.g. *Surveillance* activity corresponds to "the watching, listening to, or recording of an individual's activities" [8].
- **Threats:** Deng et al. identify a set of threats against privacy; e.g. *Information Disclosure* is the exposure of "personal information to individuals who are not supposed to have access to it" [11].
- **Vulnerabilities:** Antón et al. consider privacy as a set of vulnerabilities to be protected against [9]; e.g. *Information storage* vulnerability corresponds to the storage of data in the organisation's database.

Some works consider several representations altogether. For instance, Antón et al. [9] consider privacy both as a set of goals and as a set of vulnerabilities. Finally, to take into account this variety of representations, some approaches propose to handle privacy, not as a set of homogeneous elements (e.g. properties or threats), but as a set of heterogeneous values belonging to several **dimensions** [7], such as visibility or granularity. In this setting, each dimension may be seen as an axis composed of discrete points. For example, the extreme points of the *Visibility* dimension are *None*, "where the data should not be visible to anyone", and *All/World*, "where the data is offered to anyone with access to the data repository".

In conclusion, our short survey shows that several representations of privacy co-exist, which reflects the complexity of the privacy concept. Indeed, the definitions of key privacy concepts may subtly change depending on the context, but also across time. Therefore, comparing different approaches for privacy is not straightforward. Next section proposes a way to reduce this complexity.

4. GePyR: A new Generic Privacy Representation

This section introduces GePyR, a new generic privacy representation, as a classification scheme relying on privacy categories. It aims at reducing the complexity coming from the diversity of privacy representations, and allowing easier comparisons of solutions for privacy protection.

As explained in the previous section, we have identified seven existing representations of privacy: principles [2], [6], [10], harmful activities [8], goals [9], vulnerabilities [9], dimensions [7], threats [11], and properties [12]. GePyR subsumes all of them: Section 5 explain

how our scheme of classification can be specialized in order to recover the key concepts of the existing representations, but let us first introduce GePyR.

GePyR introduces two key generic categories, namely *Confidentiality* and *Consent*, and two additional ones which are considered less often in the literature, namely *Accountability* and *Transparency*. Even though these names are already defined in several ways in the privacy literature (e.g., [2], [10], [20]), our own definitions, introduced below, do not aim to recover any of them, since they provide large generic categories in which existing privacy representations can be classified. In particular our definitions are more general than most existing ones. For example, considering the term of *consent*, the GDPR (article 4) defines it as follows: "'consent' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;" [2]. Whereas Hoepman, in their model, defines it thusly: "consent is the obligation object" [10]. There is a wide gap between these two definitions, but they both express a facet of a generic *consent* notion. Therefore, we decide to not rely on any existing set of definitions and introduce our own definitions that are more generic than the existing ones.

Definition 1 (Confidentiality). *Confidentiality* refers to any notion related to visibility of personal data restricted to some authorized persons.

In the running example described in Section 2, this category includes any notion related to the visibility of the e-mail addresses.

Definition 2 (Consent). *Consent* refers to any notion related to the agreement between entities who will process personal data and persons who are subjects of these data.

In our running example, this category includes any notion related to the agreement between the website owner and its users, regarding the processing of users' e-mail addresses.

Definition 3 (Transparency). *Transparency* refers to any notion related to the awareness of the persons involved in personal data about their expected uses.

In our running example, this category includes any notions related to the awareness of the users about the expected uses of their e-mail addresses.

Note that awareness is necessary for users to issue a valid consent. Therefore, the *consent* category depends on the *transparency* category, according to our definitions.

Definition 4 (Accountability). *Accountability* refers to any notion related to the ability for the entities processing personal data to demonstrate that they respect the expected privacy definition.

In our running example, this category includes any notion related to the ability for the website owner to demonstrate that it respects rules related to data processing.

A demonstration required for *accountability* usually relies on a specific privacy representation. In GePyR, any privacy representation is expressed by a category. Therefore, our *accountability* category depends on our *consent*, *confidentiality*, and *transparency* categories.

GePyR only introduces these four core privacy representations in its classification scheme. In particular, even if GDPR also affords specific rights applicable for specific legal bases, we do not take them into account in this paper. We believe that adding them to GePyR and PyCO is possible, but let it to future work. However, as demonstrated in the next section, it is expressive enough to subsume most of the privacy notions in the studied papers. Since a review of the literature is never comprehensive, we do not claim that GePyR is able to express all existing or future privacy notions, but we believe that it is flexible enough to be easily extended if needed.

	Consent	Confidentiality	Transparency	Accountability
Principles	Purpose Limitation [2] [6], [10]	Integrity and Confidentiality [2]	Lawfulness, Fairness and Transparency [2]	Accountability [2]
	Storage Limitation [2]	Data Breach Notification [10]	Transparency [10]	Accountability and (Provable) Compliance [10]
	Data Minimization [2], [10]			
	Accuracy [2]			
Properties	Purpose Binding [16]	Anonymity [12]		
	Necessity of Data Collection and Processing [16]	Unlinkability [12]		
		Undetectability [12]		
		Unobservability [12]		
Goals	Choice/Consent [9]	Integrity/Security [9]	Notice/Awareness [9]	
Harmful Activities	Interrogation [8]	Identification [8]		
	Secondary Use [8]	Breach of Confidentiality [8]		
		Disclosure [8]		
		Increased Accessibility [8]		
Threats	Policy and Consent Non-Compliance [11]	Linkability [11]	Content Unawareness [11]	
		Identifiability [11]		
		Detectability [11]		
		Disclosure of Information [11]		
Vulnerabilities	Information Collection [9]	Information Aggregation [9]		
	Solicitation [9]	Information Transfer [9]		
	Information Monitoring [9]			
	Information Storage [9]			
Dimensions	Purpose [7]	Visibility [7]		
	Retention [7]	Granularity [7]		

TABLE 1. SPECIALIZATIONS OF GEPYR'S CATEGORIES.

5. Specializing our Generic Representation

The categories introduced in GePyR may be specialized in different ways. These specializations depend on the context in which the categories are used and/or the point of view from which they are considered. Table 1 shows the specializations of our categories in the existing privacy representations presented in Section 3. Cells in the table are more or less full depending on the specialization. Some categories are more present in existing representations than other categories. However, each category of GePyR can be specialized in at least one existing representation, and each existing representation is covered by at least one of our categories. Furthermore, it illustrates that our two key categories, consent and confidentiality, can be specialized in all the existing representations.

For specializing our categories in an existing representation, we refer to their respective definitions introduced in Sections 3 and 4 in order to find one existing notion for the target representation that matches our category

definition. For example, we consider that GDPR's principle of accountability, defined as "the controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1"[1] [2], specializes our accountability category, defined as "any notion related to the ability for the entities who use personal data to demonstrate that they respect the expected privacy properties.", considering this personal data processing obligation aligns with privacy concerns. Therefore, it fills the cell at the intersection of row "Principles" and line "Accountability".

Consent Specializations: Let us first consider our consent category, presented in the first column of Table 1. It can be specialized in all the privacy representations. Consider for instance the *purpose limitation* principle [2], [6], [10], which aims for personal data to be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes" [2]. This principle relies on notions related to the processing of personal data (e.g. collection) and more precisely it restricts processing according to the agreement between the entity in charge of processing personal data and interested parties. It can thus be seen as a notion related to the agreement between entities who process personal data and persons who are subjects of these data. Therefore, we consider that it may be seen as an instance of the consent category.

Another example of possible instantiation is Solove's *secondary use* harmful activity [8], since he defines it as "the use of data for purposes unrelated to the purposes for which the data was initially collected without the data subject's consent".

Confidentiality Specializations: Now, let us consider our confidentiality category, specialized in the second column in Table 1. Here, we consider as specializations all representations related to visibility of personal data restricted to some authorized entities, as the Pfitzmann and Hansen's *anonymity* property, since they define this property as "the attacker cannot sufficiently identify the subject within a set of subjects, the anonymity set" [12]. Deng et al.'s *disclosure* threat may also be included in our notion of confidentiality, since they define disclosure as the fact of "expos[ing] personal information to individuals who are not supposed to have access to it" [11].

In addition to specializing GePyR according to the existing representations, it is also possible to check that it can represent a set of particular notions of existing representations, for instance those defined by GDPR.

GDPR-based Specializations: the main GDPR principles are covered by our four categories: consent, confidentiality, transparency and accountability. These specializations into a representation as principles correspond to the first line in Table 1. Indeed, the GDPR principles *purpose limitation*, *storage limitation*, *data minimization*, and *accuracy* may be seen as specializations of our consent category, while GDPR principles *Integrity and Confidentiality* and *Data Breach Notification* may be seen as specializations of our confidentiality category. Other main GDPR principles *Lawfulness, Fairness and Transparency* and *Accountability* may be seen as specializations of our two remaining categories.

1. Paragraph 1 defines the set of principles representing privacy.

This section illustrated that our classification scheme can be specialized in several existing privacy notions. Yet, it is worth noting that confidentiality is not only a privacy concern: it is also an important notion in the closely related field of security. The other notions of our classification scheme, i.e. consent, transparency, and accountability, are unique to privacy. Therefore, we choose to focus on these privacy-specific properties in the rest of the paper.

6. PyCO: A New Privacy Context Ontology

This section introduces PyCO, our new privacy context ontology, its key elements and their relationships in the context of privacy. Verifying privacy-related properties requires to specify the verification context which is, in our case, the key privacy elements and their relationships according to the accountability, transparency, and consent categories. The elements that we consider essential for the analysis are largely inspired by the concepts defined in the GDPR(Article 4) [2] and by Working Party 29 [6]. Indeed, they define in details what is necessary to take into account when considering privacy issues, including the stakeholders, the environment and the system, but also their relationships. Furthermore, these elements are part of many existing works related to consent [15]–[20], [23]–[29].

These elements and their relationships define PyCO, our privacy context ontology, presented in Fig. 1.

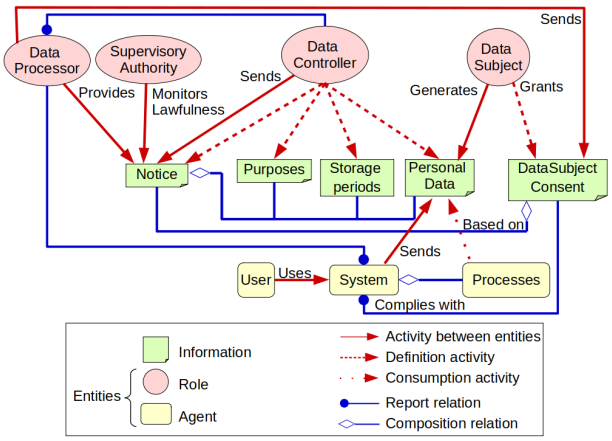

Figure 1. PyCO: Privacy Context Ontology.

It is worth noting that PyCO comes from many sources, including GDPR and more specific scientific articles, and may be seen as a common deminonator between them. Therefore, some elements and relations may slightly differ from any of these models for genericity purposes. Also, all the elements are not necessarily present in all the privacy-compliant systems. For example, in some cases, there is no (explicit) Data Processor. Below, we describe the elements included in PyCO and illustrate them with our running example, considering that the examplified R&D team uses PyCO during an implementation phase of their website lifecycle.

- **Personal Data**: data relating to an identifiable person (ex: the e-mail addresses);

- **Data Subject**: person related to personal data (ex: the website's users);
- **Data Controller**: entity that wants to process the data and, therefore, defines the purpose of its use for this processing (ex: the website's owner);
- **Data Processor**: entity that processes the data on its behalf, and provides guarantees to implement appropriate measures for privacy protection (ex: there is no explicit data processor, so the data controller, i.e. the website's owner, takes this role);
- **System and Processes**: entities that use data and compute them, respectively (ex: the functions in the code that send the subscription expiration notifications and send the targeted advertising);
- **User**: individual who uses the system (ex: the website's users);
- **Purposes**: why personal data are processed (ex: keywords "administration" and "marketing" in the source code);
- **Storage Duration**: the amount of time the data is processed (ex: some corresponding keywords, like "no-retention" and "business-practices");
- **DataSubject Consent**: the consent granted by the data subject (ex: the consent of the processing of the e-mail address for the specified purposes);
- **Notice**: set of the pieces of information necessary for establishing the consent (ex: the processed personal data, the e-mail addresses, the specified purposes, and the storage duration);
- **Supervisory Authority**: entity monitoring that everything complies with the law (ex: CNIL, the French agency in charge of protecting public and private data in digital systems).

These elements may be split into two different kinds: entities on the one hand, and information to be transferred, on the other hand. Among entities, some correspond to legal roles held by individuals, companies or organizations, depending on the context, while the others are agents, i.e. individuals or digital components in practice. In some contexts, the same person or organisation may have several roles. For instance, in our running example, the *Data Subject* and the *User* refer to the same person (the website's users), and the *Data Controller* (the website's owner) also takes the role of *Data Processor*. Indeed, the website's owner processes the personal data and is also in charge of guaranteeing that appropriate measures for data protection are implemented. Note however that there are some contexts where these entities represent different persons or organisations. For example, in a hospital software for medical record, a doctor is the *User* since he interacts with the software, but the *Data Subject* is the patient. That is why it is necessary to differentiate the *Data Subject* and the *User*.

PyCO also includes relationships between its elements. These relationships may also be split into two kinds: activities and relations. The former is an action between two entities, and may define or consume some pieces of information, while the latter are either reporting relations or composition relations. An example of information definition is the *Data Controller*, who *defines* the Notice. The relationship between the *User* and the *System* is an example of action between entities, the former using the latter. Also, *Personal Data* are consumed by the *Processes* that are part of the *System*, while the *Data Processor* processes the data on *Data Controller*'s behalf. The composition relations are based on the content of the modelled information. For example, we consider that there is a composition relation in our model, between DataSubjectConsent and Notice, because the DataSubjectConsent information is based on the information presented in the Notice. In other words, each information present in the Notice (Purposes, Personal Data and Storage Duration) has to be present in the DataSubjectConsent information.

In our running example, from the set of papers identified with GePyR, the team chooses the article of Hayati et al. [15], because it also targets website, and uses PyCO for identifiying the elements introduced in the article and how they are represented. Here, it is quite immediate from the paper that the important elements are the **Personal Data**, represented as Object in the targeted language, the **Processes**, represented as functions, the **Purposes** and the **Storage Duration**, both represented as keywords, the **Notice**, composed of the Personal Data, the Processes, the Purposes and the Storage Duration, and the **DataSubject Consent**, based on the Notice. Additionally, PyCO also allows the team to quickly and easily know which privacy-related elements are not explicitly presented and used in the selected paper. Here, they are the **Data Controller**, the **Data Subject**, the **Data Processor**, the **Users**, and the **Supervisory Authority**. In particular, thanks to this information, during the validation and verification steps of the website's lifecycle, the R&D team can easily compare their solution with the theoretical framework of Hayati et al in order to check whether the elements and their relationships are correctly represented.

7. Example of Using GePyR and PyCO

Up to now, we have defined GePyR, a new classification scheme that allows us to identify categories that include main concepts involved in privacy (and encompassing the relevant aspects of personal data protection). GePyR is more generic than existing taxonomies and, therefore, does not allow to directly propose solutions to verify privacy-related properties. For this purpose, it is necessary to specialize it according to existing representations. Nevertheless, this genericity provides a way to consider existing solutions and case studies in a uniform setting. We have also defined PyCO, a privacy-related ontology that represents the important elements and their relations that should be taken into account to handle privacy-related properties. These two contributions allow us to categorize existing privacy-focused approaches in order to simplify their comparisons (or just make them possible).

Beyond our running example that illustrates a way to use GePyR and PyCO, this section aims at demonstrating how to use them on existing works. We analyze these examples according to three different abstraction levels that correspond to different layers of a system's life-cycle:

- **H**igh-**L**evel: the abstraction level for the beginning of the life-cycle (e.g., the system requirements described within a natural language);

LVL	TARGET	REPRESENTATION		REF
HL	Location-based services	Threats	Linkability	[14]
	Communication protocols	Threats	Disclosure of information	[30]
	Trace sets	Properties	Non-interference	[31]
ML	Communication protocols	Properties	Unlinkability	[32]
	Data-flow diagram	Vulnerabilities	Information Storage	[17]
	Cyber Physical Systems	Threats	Disclosure of information	[33]
PL	Internet of Things	Principles	Data minimization	[23]
	Web privacy policies	Properties	Non-Disclosure	[15]
	Mobile applications	Principles	Data breach notification	[34]

TABLE 2. CONFIDENTIALITY-RELATED REPRESENTATIONS.

- **M**odel-**L**evel: the abstraction level for the design phase of the life-cycle (e.g., the model of the system specified within a modeling language);
- **P**rogram-**L**evel: the abstraction level for the implementation step of the life-cycle (e.g., the code that implements the system and described within a programming language).

For each classification and paper, we first identify its main category, either consent or confidentiality, based on their definitions in Section 4, and their abstraction level. Then, for each GePyR example, we identify the application domain of the paper and the specialized representation corresponding to the topic of the paper. For the PyCO instantiations, we identify the target domain and system of the paper, and how PyCO's elements are represented in the paper. Section 7.1 presents GePyR examples for classifying existing privacy approaches, while Section 7.2 presents PyCO instantiations on concrete examples from the literature.

7.1. Classification of existing privacy-focused approaches

There is a wide variety of approaches that propose solutions to ensure privacy. Each of these approaches relies on its own representation of privacy and focuses on one part of its representation. Using GePyR, we propose a unified classification for some of the existing approaches. This classification help us compare them, for instance to help detect those that address comparable privacy issues. In this section, we limit our classification on approaches related to confidentiality and consent, which are summarized in Table 2 for confidentiality and Table 3 for consent. They include three references for each abstraction level and categories: [14], [30], [31] for high level (HL) confidentiality-oriented papers, [21], [22], [25] for HL consent-oriented papers, [17], [32], [33] for model level (ML) confidentiality-oriented papers, [23], [26], [27] for ML consent-oriented papers, [15], [23], [34] for the programming level (PL) confidentiality-oriented papers, and [15], [19], [28] for PL consent-oriented papers. This choice helps compare the approaches within and/or between the different levels. The small number of papers that we compare here may introduce a bias in our study, so it prevents us from making any definitive conclusion, but it is enough for demonstrating what it is possible to do thanks to GePyR.

This classification helps to easily identify which solutions may be comparable and which ones may be complementary. For example, Table 2 shows that the AVISPA team [30] proposes a solution for disclosure of information threats for communication protocols, while Sharma et al. [33] propose a solution for the same threat but for cyber-physical social systems. Both propose a solution against the exposure of personal information to individuals who are not supposed to have access to it. They focus on the same confidentiality issue but their solutions are different: the AVISPA team [30] defines a role-based language to specify communication protocols and provides an associated tool, whereas Sharma et al. [33] provide algorithms relying on an anonymity technique and a purpose-based access control. This table also shows that several solutions ensure privacy in communication protocols but they focus on different privacy representations: either disclosure of information [30], or an unlinkability property [32]. The solution proposed by the AVISPA team is described above [30], while Hirschi et al. [32] define two conditions on protocols that are sufficient to ensure unlinkability and provide a tool for checking them. Similar observations can be made for consent through Table 3. For example, Petkovic et al. [27] propose a solution for preventing harmful activities of secondary use in hospital information systems, while Dufay et al. [28] propose a solution for the same harmful activities but applied to databases. The former proposes to use trace analysis to detect the purpose of the data uses, while the latter expresses properties in JML and uses the tool Krakatoa to check them. This table also shows that several solutions for ensuring privacy in hospital information system do exist, but focusing on different privacy representations: either harmful activities of secondary use [27], or threats of policy and consent non-compliance [19]. Tokas et al. [19] define a specification language for privacy policies and provide a tool, CASTOR, to statically check them. To sum up, this classification helps compare existing solutions and identify which privacy issues are tackled in which application domains.

Using GePyR on our Running Example: In our running example, the R&D team that develops the website may rely on GePyR in order to identify papers of interest in an existing base of knowledge. For instance, they may look for solutions that enhance consent (with our generic meaning) by following principles of privacy, including data protection aspects, in particular the GDPR principles. Assuming the use of our (arguably incomplete) state of the art as a base of knowledge, they obtain four papers [15], [21]–[23]. Then they can compare these selected solutions and choose the best with respect to their specific issue.

Besides, this classification could allow to identify patterns related to representations and categories. At the current scale, we can not form strong hypotheses, but we can already observe a few specificities that would have to be confirmed by studying more papers. A first remark arising from this classification is that some representations may be predominant for a particular category but not for the others. For instance, we observe that the majority of confidentiality-oriented approaches represent privacy either as threats, or as properties. Our hypothesis is that confidentiality is often seen as a security concern, for which this concept is studied as a key property with associated threats. For consent-related approaches, we observe that the majority of representations of privacy are principles. Our hypothesis is that it comes from the influence of GDPR. We additionally found no approach that represents privacy as dimensions [7]. We have no clear

LVL	TARGET	REPRESENTATION		REF
HL	Mobile App	Goals	Choice/Consent	[25]
	Home automation	Principles	Purpose limitation	[21]
	Web sites	Principles	Lawfulness, fairness and transparency	[22]
ML	Hospital Information System	Harmful Activities	Secondary use	[27]
	Diagnostic process	Threats	Policy and consent non-compliance	[26]
	Smart device (IOT)	Principles	Lawfulness, fairness and transparency	[23]
PL	Hospital Information System	Threats	Policy and consent non-compliance	[19]
	Web sites	Principles	Purpose limitation	[15]
	Database	Harmful Activities	Secondary use	[28]

TABLE 3. CONSENT-RELATED REPRESENTATIONS.

LVL	DOMAIN	SYSTEM	PROCESSES	PURPOSES	PRIVATE DATA	STORAGE PERIODS	REF
HL	Public Transport	Mobile App	DNL*	Keywords	DNL*	∅	[25]
	Home Automation	Vocal Assistants	DNL*	DNL*	DNL*	∅	[21]
	Web Services	Web Sites	DNL*	DNL*	DNL*	∅	[22]
ML	Medical	IT	BPM**	Keywords	Keywords	∅	[27]
	Medical	Diagnostic Process	Markov Decision Process	Decision function	Keywords	∅	[26]
	Smart Building	Smart device (IOT)	BPM**	Keywords	Keywords	∅	[23]
PL	Medical	IT	Functions	Keywords	Keywords	∅	[19]
	Web services	Web sites	Functions	DNL* or Keywords	"Object"	Keywords	[15]
	Human resources	Database	Functions	Keywords	"Objects"	∅	[28]

* Descriptions in Natural Language
** Business Process Models

TABLE 4. PYCO INSTANTIATION ON CONCRETE EXAMPLES.

hypothesis for this observation. It might come from the non-exhaustiveness of our study, or because it would be of a different nature than the other existing representations. We postpone to future works further investigations about the three hypotheses that are raised by our classification.

7.2. Instantiation of PyCO on examples

This section demonstrates how our ontology PyCO can be instantiated on existing solutions that verify consent properties. Indeed, Table 4 shows how some PyCO elements can be instantiated on concrete use cases on which these solutions have been applied. More precisely, it includes the most significant elements for comparing different context representations in these approaches:

- the abstraction level;
- the application domain;
- the type of system;
- the key modeled elements: processes, purposes, personal data and storage period.

It does not include other elements such as users, data subjects, data controllers, data processors, and supervisory authorities because they are not a differentiating factor: when they are considered, they are modeled in the same way (for instance, always as a person or always as a company), regardless of the application domain or the abstraction level. Additionally, it does not include the notice and the data subject consent, since the only differences actually come from other elements already shown in Table 4.

To illustrate how PyCO can be instantiated on various examples, we have chosen three references for each abstraction level: [21], [22], [25] for the high level (**HL**), [23], [26], [27] for the model level (**ML**), and [15], [19], [28] for the programming level (**PL**). This choice makes it easier to identify the differences and similarities within and/or between the different levels.

Some of these approaches, such as [15], [22] for web services, have the same application domain, even if they study it at different levels of abstraction. In this case, it is particularly interesting to identify how the important elements to verify privacy-related properties are modeled in these approaches and to compare them. For example, in the approaches for web services, the processes may be described either in natural language [22], or by functions of some programming language [15].

Furthermore, thanks to our context modeling, we can notice similarities within the same level of abstraction. For example, at **PL**, processes are always represented as functions in all the considered examples. Such an observation may be used in several ways. First, this can help compare examples from the literature. Second, it can help identify formalisms already used at different levels, in order to propose a new solution compatible and/or comparable with existing works. In contrast, it could also be used to identify that some research directions have not yet been investigated (or, more precisely, have not yet produced any published results).

Table 4 also makes it possible to identify elements that have rarely been considered in existing works. For example, from our state of the art, the supervisory authority is only considered in the context of the CNIL judgment against Google [22]. In the same way, only Hayati and Abadi [15] consider the storage periods in their privacy-related property verification solution.

7.3. Use of our Contributions on the Running Example

This section illustrates a possible methodology for using our contributions on the running example presented in Section 2. As a reminder, consider an R&D engineer team in a French organization that wants to develop a new website which relies on some user personal data for two purposes: website administration on one hand and marketing on another hand. Among others, the website uses the users's e-mail addresses to inform them when their subscription expires, but also to send them targeted advertising. Since it is used in Europe, the system need to be GDPR compliant. For that purpose, during the lifecycle of the system, the team apply the methodology of Figure 2, based on our contributions GePyR and PyCO. We only present here the global scheme of the methodology. How to concretely use GePyR and PyCO on the presented system is illustrated later in the paper.

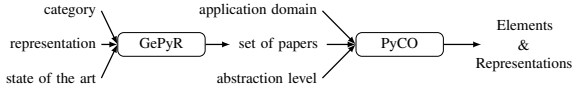

Figure 2. One Possible Methodology for Using GePyR and PyCO.

If we assume a spiral model [5] for implementing the system, GePyR and PyCO are used for helping designing and implementing the website during the first iteration, and also used for verifying the system in the following

iterations. More precisely, the team first uses GePyR, our classification scheme detailed in Section 4, to identify existing solutions to their privacy concerns. To this end, they first choose a category on which they focus, e.g. the Consent category, and the representation matching their point of view: since the website is deployed in Europe, the team must follow GDPR data protection principles rules [2] (which will impact the level of privacy offered by the service). From these inputs, and thanks to GePyR, a set of articles proposing solutions to the corresponding issues can be identified in the state of the art. The team can then select a few papers based on their abstraction level and/or their targeted application domain, and use PyCO, our ontology detailed in Section 6 to identify which privacy elements are useful, how to represent them, and which relations between them should be implemented. During future iterations, PyCOhelps check whether the system indeed implements these elements and their expected relations during the validation and verification steps.

To sum up, even if our state of the art is not exhaustive, it looks like the classification allowed by our modeling helps unify representations of privacy proposed in existing approaches but applied in different contexts. This way, it could be possible to identify similarities, but also differences. It seems also possible to detect less studied elements, as well as combinations of categories that have not been published. It should be validated by extending our study to additional references.

8. Related Work

Several ontologies have been proposed for dealing with privacy. Gharib et al. [35] provide an implemented ontology for privacy requirements from laws based on a systematic literature review. It aims at providing a set of concepts to analyze privacy requirements in their social and organizational context.

Pandit et al. [36] define an ontological design pattern that describes personal data in privacy policies to build a common representation for activities using information contained in privacy policies, in particular for sharing this information.

Additionally, Loukil [37] proposes a privacy-based approach for ensuring privacy in IoT according to laws and regulations. In this approach, an ontology is defined to describe the IoT environment and privacy requirements.

Besides, Oltramari et al. [38] define a semantic framework for analysis of privacy policies. They aim at improving the understanding of privacy policies for data subjects and supporting research analyses of them. They implement a tool based on this semantic framework. Even if they do not define an ontology for modelling all privacy concepts, they rely on concepts linked to privacy and to our ontology.

Three of these ontologies [35], [37], [38] do not follow the same methodology as us. They start from the requirements, or from the privacy policies to deduce an ontology, because their goals are different from ours. We also notice that, although one of them [37] is specialized for IoT, the others are generic, like ours.

Pandit et al.'s ontology [36] is closest to ours, since it follows a similar methodology: starting from the system and from existing solutions from the literature in order to establish the ontology. We share the same goal and both ontologies have many common elements. However, there are several differences. First, their approach does not distinguish between processes and purposes. Therefore, it is for instance not adapted when studying the sentence of Google by CNIL [22], since the CNIL ruled that the lack of transparency of Google was due to asking consent for each service (i.e. processes), and not for specific purposes: distinguishing processes and purposes in that case is necessary for establishing that the consent was not "specific" and "unambiguous". Another difference is that they predefine some process types, while we do not. However, we believe that, thanks to its generic side, our ontology can represent these specific processes. Last, Pandit et al. focus on what is needed for privacy policies, while we consider a broader context, including for example the various actors involved in privacy. Some approches propose surveys on existing legal ontologies. Kurteva et al. propose a survey of ontologies focusing on consent as defined in GDPR [39]. They provide a detailed overview presenting the consent specificities and they legals aspects. Along this overview, they define "best practices" for ensuring compliance on consent. Leone et al. provide a structured comparative analysis for recent legal ontologies [40]. They compare in detail these ontologies on specific criteria like used languages or adopted norm models. However, both only rely on legal aspect. In particular, none of them highlights common elements of privacy represented in these ontologies. On our side, we propose a more generic representation of privacy, even if it is less detailed and less specific. It allows us to highlight key privacy elements, and their relationships, relying of various sources such as GDPR and existing taxonomies [7]–[9], as explained in Section 3.

9. Conclusion and Future Work

This paper proposes the new privacy representation GePyR, a generic and specializable classification scheme relying on privacy categories. It is generic enough to take into account various contexts and different levels of abstraction, and it can be adapted to the possible evolution of existing representations. Specializing GePyR for the existing representations demonstrates that it subsumes them. It also allows to compare existing solutions enhancing privacy protection. We illustrated on concrete examples how it can be specialized.

However, this classification scheme does not introduce several key privacy elements. Therefore, we also propose PyCO, an ontology for privacy that describes them and their relationships, with a focus on consent. PyCO helps to compare existing approaches related to privacy, in particular the differences in their representations of privacy elements.

Future work includes extending our state of the art to a larger set of papers that would further demonstrate the usefulness of GePyR and PyCO. We have also planned various steps arising from these contributions to help verify compliance with privacy categories. From our model, we plan to identify a set of properties related to our category of consent that would be relevant for privacy. Our model will also allow us to identify the different key elements to take into account for verifying these

properties at different abstraction levels. As mentionned in the introduction, in this paper we used our own graphical representations for our contributions to represent all of our concepts in a visual, clear and succinct way. A future work could consist in translating our diagram in various UML diagrams to represent all our concepts in order to allow a more detailled reading.

References

[1] L. Brandeis, and S. Warren, "The right to privacy," 1989.

[2] European Commission, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[3] Australian Government, "The Privacy Act," 1988. https://www.oaic.gov.au/privacy/the-privacy-act/

[4] Personal Information Protection Commission, "Act on the Protection of Personal Information," 2016. https://www.ppc.go.jp/files/pdf/Act_on_the_Protection_of_Personal_Information.pdf

[5] B. Boehm, "A spiral model of software development and enhancement," 1988.

[6] Working Party 29, "ARTICLE 29 DATA PROTECTION WORKING PARTY - Opinion 03/2013 on purpose limitation," 2013.

[7] K. Barker, et al., "A data privacy taxonomy," 2009.

[8] D. Solove, "A taxonomy of privacy," 2005.

[9] A. Anton, and J. Earp, "A requirements taxonomy for reducing web site privacy vulnerabilities," 2004.

[10] J. Hoepman, "Privacy design strategies," 2014.

[11] M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen, "A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements," 2011.

[12] A. Pfitzmann, and M. Hansen, "A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management," 2010.

[13] S. Delaune, M. Ryan, and B. Smyth, "Automatic verification of privacy properties in the applied pi calculus," 2008.

[14] P. Asuquo et al., "Security and privacy in location-based services for vehicular and mobile communications: an overview, challenges, and countermeasures," 2018.

[15] K. Hayati, and M. Abadi, "Language-based enforcement of privacy policies," 2004.

[16] S. Fischer-Hubner, and A. Ott, "From a formal privacy model to its implementation," 1998.

[17] I. Oliver, "Privacy engineering: A dataflow and ontological approach," 2014.

[18] P. Ashley, S. Hada, G. Karjoth, C. Powers, and M. Schunter, "Enterprise privacy authorization language (EPAL)," 2003.

[19] S. Tokas, O. Owe, and T. Ramezanifarkhani, "Language-based mechanisms for privacy-by-design," 2019.

[20] X. Zhang, F. Parisi-Presicce, R. Sandhu, and J. Park, "Formal model and policy specification of usage control," 2005.

[21] CNIL, "Exploration des enjeux éthiques, techniques et juridiques des assistants vocaux," 2020.

[22] N. Certes, "RGPD : Google condamné à 50 millions d'euro par la CNIL," 2020. https://www.lemondeinformatique.fr/actualites/lire-rgpd-google-condamne-a-50-meteuro-par-la-cnil-74062.html

[23] M. Barati, O. Rana, I. Petri, and G. Theodorakopoulos, "GDPR Compliance Verification in Internet of Things," 2020.

[24] J. Laakkonen, S. Annala, and P. Jappinen, "Abstracted architecture for smart grid privacy analysis," 2013.

[25] M. Robol, T. Breaux, E. Paja, and P. Giorgini, "Consent Verification Under Evolving Privacy Policies," 2019.

[26] M. Tschantz, A. Datta, and J. Wing, "Formalizing and enforcing purpose restrictions in privacy policies," 2012.

[27] M. Petkovic, D. Prandi, and N. Zannone, "Purpose control: Did you process the data for the intended purpose?," 2011.

[28] G. Dufay, A. Felty, and S. Matwin, "Privacy-sensitive information flow with JML," 2005.

[29] A. Ahmadian, "Model-based privacy by design," 2020.

[30] The AVISPA team, "HLPSL Tutorial," 2006.

[31] M. Clarkson, and F. Schneider, "Hyperproperties," 2010.

[32] L. Hirschi, D. Baelde, and S. Delaune, "A method for unbounded verification of privacy-type properties," 2019.

[33] T. Sharma, J. Bambenek, and M. Bashir, "Preserving Privacy in Cyber-physical-social Systems: An Anonymity and Access Control Approach," 2020.

[34] D. Sun, C. Guo, D. Zhu, and W. Feng, "Secure HybridApp: A detection method on the risk of privacy leakage in HTML5 hybrid applications based on dynamic taint tracking," 2016.

[35] M. Gharib, P. Giorgini, and J. Mylopoulos, "COPri v. 2—A core ontology for privacy requirements," 2021.

[36] H. Pandit, D. O'Sullivan, and D. Lewis, "An Ontology Design Pattern for Describing Personal Data in Privacy Policies," 2018.

[37] F. Loukil, "Towards a new data privacy-based approach for IoT," 2019.

[38] A. Oltramari et al., "PrivOnto: A semantic framework for the analysis of privacy policies," 2018.

[39] A.Kurteva, T.Chhetri, H.Pandit, and A.Fensel, "Consent through the lens of semantics: State of the art survey and best practices," 2021.

[40] V.Leone, L.Di Caro, and S.Villata, "Taking stock of legal ontologies: a feature-based comparative analysis," 2020.