





P2Onto: Making Privacy Policies Transparent

Evgenia Novikova^{1,2} , Elena Doynikova^{1,2}  , and Igor Kotenko^{1,2} 

¹ HUAWEI Research Center, Marata str. 69-71, St. Petersburg, Russia
{novikova, doynikova, ivkote}@comsec.spb.ru

² SPC RAS, SPIIRAS, 39, 14 Line, St. Petersburg 199178, Russia

Abstract. The privacy issue is highly relevant for modern information systems. Both particular users and organizations usually do not understand risks related with personal data processing. The ways an organization gathers, uses, discloses, and manages a customer's or client's data should be described by privacy policy, but in major cases such policies are confusing for the customer. The goal of this research is making privacy policy transparent for the users via automation of the privacy risks assessment process based on the privacy policy. The paper introduces the developed common approach to privacy risks assessment based on analysis of privacy policies and ontology for privacy policies. The approach includes construction of an ontology for a privacy policy, and generation of rules for privacy risks assessment based on the proposed ontology. The applicability of the proposed approach and ontology is demonstrated on the case study for IoT device.

Keywords: Privacy policy · Privacy risks · Personal data · Ontology · Semantic analysis · Natural language processing · Risk assessment

1 Introduction

The privacy issue is not novel for modern society. From the moment the various processes began moving into the information space, a large amount of personal information moved there. Personal data is data that identifiably describe a living individual person [1]. This information may be of financial interest, and it is used by different companies for a variety of purposes. In addition to using with so-called legal purposes, i.e. the purposes that have legal basis, this information can be stolen if information security requirements are not satisfied. It should be noticed that though the privacy issue is under discussion for the many years, the individuals, generally, do not understand what is personal data, how and when they provide legal basis for using their personal data while interacting with systems, products, and services, how and when their personal data can be stolen, as well as how personal data can be used against the individuals (e.g. annoying advertising, black PR (Public Relations), black market, damage to reputation, etc.). At the same time, the organizations that provide the systems, products, and services, may not completely realize the consequences of the personal data leakage both for their customers, and the organizations themselves. These consequences can include the financial losses, damage to reputation, and negative impact for the organization's development.

A number of incidents involving the personal data leakage led to the development of the EU General Data Protection Regulation (GDPR) [1] that emphasizes control over personal data and states that data subjects should be made aware of the risks related to personal data processing. This has forced the organizations to pay more attention to the privacy issues to avoid law and financial problems. The organizations should generate a privacy policy while providing various information products. Privacy policy is a statement or a legal document (in privacy law) that discloses some or all of the ways a party gathers, uses, discloses, and manages a customer or client's data. But most users accept such policies without even reading and understanding what kind of data and on what period they provide. A representative example is the application developed in Moscow to track the movements of individuals infected with COVID. All individuals that install the application and accept the privacy policy, give their consent to transfer all the data that application can get (from IP address to the passport ID and employer) to any third parties for almost any purposes including advertising for 10 years.

The goal of this research is making privacy policy transparent for the users by automation of the privacy risks assessment process based on the privacy policy. The approach based on the ontology is proposed.

As it is mentioned, the analysis of privacy policies is highly relevant and not novel issue. But to this moment there is no completed research related to the risk assessment based on the privacy policies analysis.

In this paper we propose an approach that incorporates analysis of the privacy policies written in natural language for the subsequent formal specification of the policies using an ontology and privacy risk assessment based on the constructed ontology.

The main contributions of the paper are as follows:

- a common approach to privacy risks assessment based on ontology constructed for a privacy policy;
- a privacy policy based ontology;
- an approach for constructing rules for privacy risks assessment based on the proposed ontology;
- a usage scenario.

The paper is organized as follows. Section 2 analyzes the main related works in the area of formal languages development for the privacy policies specification, application of natural language processing for the privacy policies analysis, existing privacy aware ontologies and privacy risks assessment. In Sect. 3 the developed methodology for privacy risks assessment is introduced, the developed ontology is provided, including the design and implementation processes, key concepts and application (Subsect. 3.1), converting privacy policy text to ontology (Subsect. 3.2), and privacy risks assessment procedure (Subsect. 3.3). In Sect. 4 the case study on application of the developed ontology to assess privacy risks is given, examples of the rules that can be used for privacy risks assessment are considered, and the discussion on advantages of the suggested ontology and the proposed privacy risks assessment procedure is provided. The paper ends with conclusion and future research directions.

2 Related Works

In this section we outline and analyze two main groups of researches related with ours.

The first group of the researches covers development of formal languages. The proposed languages can be used to specify the policies, while the developed policies can be used for the further analysis or privacy risk assessment. We consider this group of works as soon as in the scope of our approach we should develop a formal language for subsequent ontology specification.

The second group of the researches covers analysis of privacy policy texts represented using natural language. We consider this group of works as soon as privacy policies are usually generated using natural language. To develop formal language for ontology specification automatically we analyze the text of the privacy policy given in natural language first. In scope of this group of works we consider both the papers devoted to the natural language processing (NLP) and the papers devoted to the NLP application to analyze privacy policies and to assess the risks.

The formal languages are used for specification of the security policies, license agreements, access control policies, and privacy policies. The essence of approaches devoted to development of formal languages is specification of the language alphabet and of the rules for constructing the sequences using the characters of the alphabet (i.e. the language grammar). The text specified using such language can be processed using mathematical methods. There are a lot of application areas of this approach. As soon as this research is devoted to the privacy policies processing, we review in details the papers that consider development of formal languages for specification of privacy policies.

In [2] the authors propose the Platform for Enterprise Privacy Practices (E-P3P) to formalize a privacy policy into a machine-readable language that can be enforced automatically within the enterprise by the means of an authorization engine. The formalized policy specifies what types of the personally identifiable information (PII), for what purposes and by what users in the organization can be used. To formalize the policy the language that incorporates the terminology and the set of authorization rules is used. The terminology includes six elements, namely, data categories, purposes, data users, the set of actions, the set of obligations and the set of conditions. The authorization rules are used to allow or deny an action. Similar approach to authorization management and access control is introduced in [3]. The proposed model consists of users/groups, the accessed data, the purposes of access, and access modes. It is used to ensure that personal information is used only for authorization. Authors also proposed a privacy language based on the proposed model. This language is used for privacy and access control rules formalization and automated enforcement of these rules by the means of the access control system. The proposed model is limited only by the access control considering privacy aspects.

In [4] the language based approach is also used. The authors consider the privacy principle that states that the user's personal data can't be used for the purpose different from the one that they were collected for without consent of the concerned user. The authors assume that in major cases the users do not have any idea how and what purposes their personal information is used for. To resolve this issue the authors propose a data handling policy (DHP) showing users who and under what conditions can process their personal data. This policy can be developed by the service provider or by the user using

the developed DHP language. The language incorporates the set of terms (namely, recipients, actions, purposes, PII, conditions, provisions and obligations) and rules. The DHP then enforced using policy decision points (make decision regarding the access request) and policy enforcement points (implement decision) of the access control system. The disadvantage is that such policy should be developed for each new product.

In [5] another language called PILOT for privacy policy specification is proposed. The authors also developed a tool that allows assessing privacy related risks if the policy is specified using the proposed language. The advantage of the approach is that it allows assessing the risks. The disadvantage is that this approach doesn't allow assessing them automatically if the policy is not specified using the developed formal language. The authors propose to users define the privacy policies themselves and then represent the risks of the developed policy. It is also not clear from the article how to define all possible risks that are required to get assessment for the specific risk.

In [6] the authors proposed the Layered Privacy Language (LPL) that fulfill the introduced requirements, namely, differentiation between the source and recipient of data, generating privacy policies considering the purposes of operations with data, guarantee of human-readability based on layering of privacy policies. The disadvantages of this work are as follows: the research is not completed and the proposed language does not cover all privacy aspects now; the company should define their privacy policy using LPL before analyzing it.

The privacy risks assessment approach is proposed in [7]. It is based on the harm trees. The trees are constructed based on information about the system, the personal data, the relevant risk sources, the relevant events and their impacts on privacy. The harm tree nodes are represented as triples incorporating personal data, system component, and risk source. The root node of the harm tree corresponds to a privacy harm. The leaf nodes correspond to the exploitation of data by the most likely risk source. The users' privacy settings are also considered while calculating the likelihood of the privacy harms.

The main difference between the papers of this group and our approach is that we propose generating and processing the ontology automatically for every policy specified in natural language using NLP.

The second group of the researches covers analysis of texts written in natural languages, including privacy policy texts. In [8] authors presented a pipeline for automatic privacy policy extraction and analysis of the Android applications. The main contribution is annotated corpus of the privacy policies APP-350 Corpus available by link: <https://www.usableprivacy.org>. The authors applied the TF-IDF (term frequency and inverse document frequency) approach to construct feature vector from text of the policies and the support vector machine (SVC) classifier to detect different data practices in policies. In [9] the authors applied machine learning approach to automated detection of opt-in/opt-out choices to control personal data visibility. They tested different machine learning techniques for policy text analysis, such as linear regression and neural networks and experimented with different set of features. However, application of the approach requires labeled data set. The authors implemented this procedure manually.

In [10] the authors propose a semantic framework PrivOnto to analyze privacy policies. The proposed framework uses as input the set of annotated privacy policies and developed an ontology representing a set of policies with identified privacy aware data

practices. The key challenge here was related to the automated annotation of privacy policies to generate specific ontologies, for this goal crowdsourcing, machine learning and natural language processing were used. First, the experts analyzed the set of privacy policies and annotated them using outlined 11 categories of data practices (First Party Collection/Use:Privacy, Third Party Sharing/Collection:Privacy, User Choice/Control, User Access, Edit, & Deletion, Data Retention, Data Security, Policy Change, Do Not Track, International & Specific Audiences, Other). These categories served as main concepts to model privacy policies. Annotated set then was used to train the framework for automated annotation. The researchers annotated over 23,000 data practices extracted from 115 privacy policies and made them publicly available by link: <https://www.usableprivacy.org>. This research is the most closest to ours, but we focus not only on detection of data practices in text of policies, but on assessment of risks for personal data. To achieve this we focus on a particular privacy policy and develop a detailed semantic presentation of each privacy-aware data practice.

In this paper we introduce the proposed ontology that is the basis for our approach to the automated analysis of privacy risks based on the privacy policies. Though some aspects related to privacy policies analysis are covered in the related research, today there is no end-to-end approach to automated privacy risks assessment based on policies. Thus, the main contribution of our research is a new approach to privacy risks assessment based on analysis of privacy policies defined using natural language and ontology for privacy policies.

3 Methodology

In Fig. 1 the suggested risk assessment procedure based on analysis of privacy policies is shown.

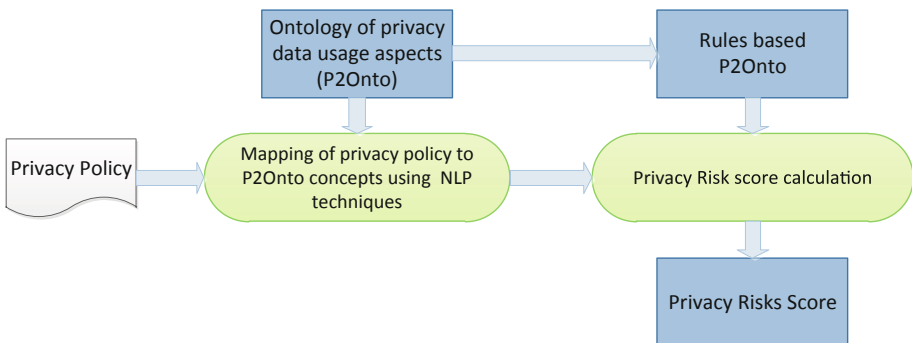


Fig. 1. General scheme of privacy risks calculation based on privacy policy analysis

The key element of the suggested approach is the P2Onto ontology that describes different aspects of personal data processing such as first party collection, third party sharing, etc. It serves as the basis for constructing an ontology for each particular privacy policy. The mapping of individuals to its concepts is implemented using natural language

techniques. P2Onto ontology also serves as the basis for constructing rules for automated privacy risk calculation. All these steps with particular focus on P2Onto ontology are described below in detail.

3.1 P2Onto Ontology

The goal of P2Onto ontology is to describe possible data usage scenarios that involve personal data processing, and to provide formal basis for the risk assessment. To construct the ontology, we used the data usage practices and associated privacy aspects proposed in [10]. These aspects were identified by the domain experts who studied both existing privacy policies and corresponding legal regulations and requirements, such as COPPA [11], and the HIPAA Privacy Rule [12]. They are listed below.

First-Party Data Collection and Usage. This aspect characterizes what personal data are collected by the service provider, operating the device, web site or application, how they are collected, what legal basis and purposes of data collection are.

Third-Party Data Collection and Sharing. This aspect characterizes all issues concerning data sharing procedures, including form of data shared – aggregated, anonymized or raw.

Data Security. This aspect describes security mechanisms, both technical and organizational, used to protect data.

Data Retention. This aspect characterizes temporal issues of personal data processing and storage.

Data Aggregation. This aspect defines if service provider aggregate personal data.

Privacy Settings. This practice defines available tools and options to end user to limit scope of personal data being collected (opt-in/opt-out issues of personal data collection).

Data Control. This aspect relates to tools and mechanisms provided to user to manipulate with personal data – access, edit, and erase.

Privacy Breach Notification. This aspect relates to the tools and mechanisms the service provider uses to inform about breach of personal data privacy.

Policy Change. This aspect relates to what tools and mechanisms the service provider uses to inform an end user about changes in text of personal data privacy and possible reactions available to end user.

Do Not Track. This practice describes how tracking signals for online tracking and advertising are processed.

International and Special Audience. This aspect discusses different issues relating with processing personal data of special audience such as children, and citizens of certain states and regions.

According to the workflow for designing ontologies based on privacy policies proposed in [13], the definition of the competence questions for each privacy aspect is a key issue that specifies the goal and tasks of the ontology. We used this approach for constructing the P2Onto ontology and determined a set of competence questions specifying issues associated with them. These competence questions are based on guidelines and questionnaires provided by international security IoT assessment frameworks such as IoTF, GSMA in the field of privacy risk assessment [14, 15]. The examples of competence questions for some privacy aspects are given in Table 1.

Table 1. Some privacy aspects and corresponding competence questions

Privacy aspect	Competence questions	Examples
First-party data collection and usage	What data categories are collected?	Geo location, activity tracking, health status, financial info, contact info, etc.
	What is the data collection mode?	Automatically without user consent, automatically but with given consent every time when automatic collection performed, or given by user directly (i.e. financial data)
	What is the purpose of data collection?	Service provision including additional services, enhancement of service provision, analytics and research, marketing and advertising, personalization, security and support services, legal requirement, etc.
	What is the basis for data collection	User given consent, legal requirement, other
	Do you collect data from third party service providers?	No, public sources, third-party service providers, others
Privacy settings	Who provides privacy settings control?	First-party service provider, Third-party service provider (including web-browser privacy settings)
	How are they implemented?	Opt-in (user directly specifies what data to collect and share), opt-out using web-link or mailing, stop using services

These questions helped us to identify core concepts and properties of the P2Onto ontology. We outlined four core concepts – *Data*, *Activity*, *Agent* and *Mechanism* – that serve as the basis for describing all aspects of data processing including tools involved in this process. Let us consider them in detail.

Data is a generic concept, it is a super class for *Personal_Data* and *Non-Personal_Data*. The concept *Personal_Data* is defined in GDPR text as “any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person” [1, Article 4].

This allowed us to determine *Sensitive_Data* concept and its subclasses to describe racial or ethnic origin (*Racial_Data*), political opinions, religious or philosophical beliefs (*Religion_Data*), genetic data (*Genetic_Data*), biometric data for the purpose of uniquely identifying a natural person (*Biometric_Data*), data concerning person health (*Health_data*), data about crime records (*Crime_Data*). We also outlined *Tracking_Data* concept to have possibility to answer Do Not Track data usage aspect. Concept *Non-Personal_Data* is used to describe non-personal data such as statistical data and is valuable to understand how many types of data – identifiable and not – are collected about particular device user.

Figure 2 shows the hierarchy of *Data* subclasses.

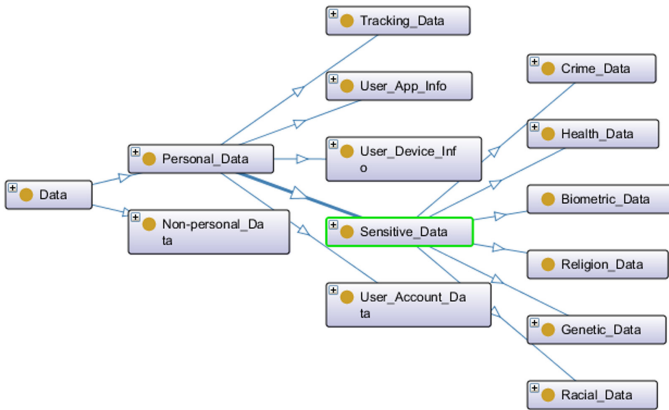


Fig. 2. Structure of data concept

Concept *Activity* (Fig. 3) is a generic concept that may be used to describe possible actions concerning data processing and data control activities. That is why we defined two different subclasses *Data_Activity* and *Control_Activity*. The first subclass is purposed to describe possible activities arising with data processing – collection, usage (or processing), storage and sharing with third parties, while the purpose of the second subclass is describe wide variety of activities associated with data privacy control and data access operations available to user, consent giving and withdrawal. It also includes

activities of service provider concerning notifications in case of policy change and breach of data. Each individual or subclass of *Data_Activity* concept has property *hasLegalBasis* that defines legal basis for data activity, including data collection. The legal basis is represented by a concept *Legal_Basis*. The purpose of data activity is described by *Data_Activity_Purpose* concept. We assume that there is a variety of data processing purposes but in general case they may fall into categories listed in Table 1.

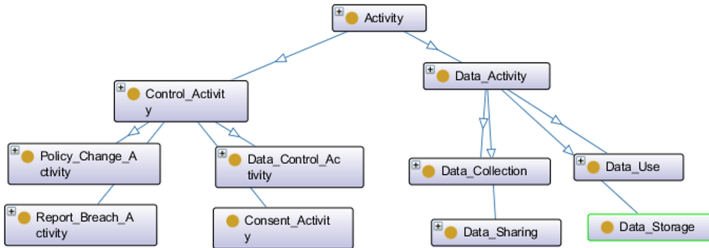


Fig. 3. Structure of activity concept

Two other important concepts are *Agent* and *Mechanism*. The concept *Agent* is used to describe service provider, end user, i.e. data subject, and third party participating in data processing. We currently suggest reusing this concept from PROV-O ontology that specifies a concept *Agent* as a subject that “bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent’s activity” [16].

The *Mechanism* class is a generic class that is used to define different mechanisms, tools or interfaces for implementing different types of data activities. It serves as a superclass to describe tools and mechanisms to collect and share data, options available to user to access data and control their privacy. The *Mechanism* is a superclass for the *Notification_Mechanism* concept used to describe ways the server provider notifies the data subject in case of data breach or privacy policy change. These classes are linked to data subjects or activities using special object properties reflecting the relationship between corresponding classes. For example, to describe security mechanisms and tools used to secure data processing, we use the property *isSecuredBy* linking the *Data_Activity* concept with the *Security_Mechanism* concept.

Figure 4 shows main concepts and properties related to the First Party Collection aspect.

3.2 Mapping of the Policy Text into P2Onto Concepts

Mapping text of the privacy policy into the concepts of the P2Onto ontology is a critical process. In major cases, text policies are monolithic texts structured in paragraphs. Some policies present important information in the form of tables or lists. This allows us to make following assumptions:

- if policy is a text organized in paragraphs, then each paragraph represents a set of P2Onto concepts semantically linked to one privacy aspect and data usage scenario;

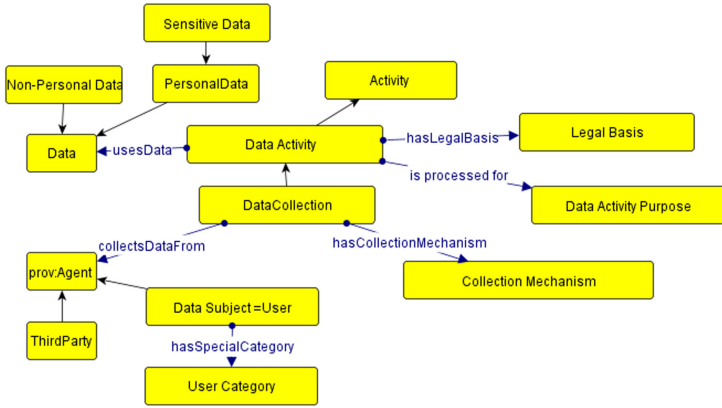


Fig. 4. P2Onto concepts and properties describing first party collection practice

- if paragraph contains a list, then each list item represent individuals of one concept or a set of P2Onto concepts semantically linked to one privacy aspect;
- if text contains a table then each row is treated as one data usage scenario, where columns contain individuals relating to different P2Onto concepts;
- if policy text is monolithic and does not contain paragraphs, then we treat it as one usage scenario, detecting individuals relating to P2Onto concepts, without linking them to one scenario.

P2Onto concepts are instantiated by words or phrases from a text policy based on simple matching them to a vocabulary that contains key words for each P2Onto concept, extended by generated synonyms.

As the result mapping each P2Onto concept will be assigned a set of individuals, if the P2Onto concept except the *Personal_Data* class and its subclasses was not detected in a given data usage scenario, then it is assigned the *NotDef* individual.

3.3 Privacy Rule Construction

The P2Onto ontology is constructed in such a way that one concept or subset of concepts may provide an answer to one competence question. Table 2 shows some examples of mapping between privacy aspects, competency questions and P2Onto concepts and properties.

Mapping one concept to one competency question allows us to propose the following privacy risk assessment procedure.

Let *PA* is a particular privacy aspect, and it includes *n* competence questions. Let *CQ_i* is *i*th competence question, and then the risk score for *PA* privacy aspect is defined as follows.

Table 2. Privacy aspects and P2Onto concepts and properties

Privacy aspect	Competency questions	P2Onto concepts	P2Onto properties
First party collection/use	What data categories are collected?	Data and its subclasses (personal data, non-personal data)	usesData
	Do you collect data from third party service providers?	Third party	CollectsDataFrom
	What is the data collection mode?	Collection mechanism	hasCollectionMechanism
	What is the purpose of data collection?	Data activity purpose	isProcessedFor
	What is the legal basis of collection?	Legal basis	hasLegalBasis
Privacy settings	Who provides privacy settings control?	Agent	providedBy
	How are they implemented?	User_Control_mechanism	Implements
	What data types do they affect?	Data and its subclasses (personal data, non-personal data)	Involves

1. For each competence question CQ_i
 - a) define a C_i concept or a set \mathbf{C}_i of concepts (belonging to one superclass, i.e. class *Data*),
 - b) calculate $RiskScore(CQ_i)$ risk score for competence question CQ_i as a $RiskScore(C_i)$ risk score of instances belonging to C_i concept, i.e. $RiskScore(CQ_i) = RiskScore(C_i)$. $RiskScore(\mathbf{C})$ for a set of concepts is defined as follows $RiskScore(\mathbf{C}) = \max\{RiskScore(C_i), C_i \in \mathbf{C}\}$.
2. Calculate privacy risk score for privacy aspect as a sum of risk scores for each competence question CQ_i :

$$RiskScore_{PA} = \sum_{i=1}^n RiskScore(CQ_i). \quad (1)$$

3. If $RiskScore_{PA} \geq High_Threshold$, then privacy risks are High, if $RiskScore_{PA} < Low_Threshold$, then privacy risks are Low, else they are Medium.

The values of threshold need to be determined during experiments after some statistical distribution of risks is obtained, but currently we suggest defining them as follows:

- High_Threshold = $4/3 \cdot n$,
- Low_Threshold = $2/3 \cdot n$,

where n is the number of competence questions defined for privacy aspect PA . The overall privacy risks are calculated as sum of $RiskScore_{PA_i}$ determined for each privacy aspect PA_i .

To calculate $RiskScore(C)$ based on individuals of the concept C , we propose to rank them as *critical*, *generic* and *other*. The rank of individuals is determined for each concept individually. Let us consider the following example, the purposes of the data collection may be as follows: p1 – service provision including additional services, p2 – enhancement of service provision, p3 – analytics and research, p4 – marketing and advertising, p5 – personalization, p6 – security and support services, p7 – legal requirement. The purposes p1 and p2 are rather generic, it is rather difficult to judge whether the data collected are really necessary or not, we propose to rank them as *generic*; purposes p3, p4 and p5 assume data aggregation and possible user profiling that is why we suggest ranking them as *critical*, purposes p6 and p7 are clear and we rank them as *other*.

Let us define following functions:

- $Critical(C)$ returns a number of individuals of the concept C that have critical rank;
- $Generic(C)$ returns a number of individuals of the concept C that have generic rank;
- $Other(C)$ returns a number of individuals of the concept C that have other rank;
- $Not_defined(C)$ returns a number of *NotDef* individuals assigned to the concept C .

Then in general case we propose using the following rule to score the risks for each concept C :

If $Critical(C) > 0$, then $RiskScore(C) = 2$, else

If $Others(C) = 0$ or $\frac{Generic(C)+Not_defined(C)}{Others(C)} \geq 1$, then $RiskScore(C) = 1$, else $RiskScore(C) = 0$.

However, in some cases it is necessary to define individual rules for some concepts. The example of such concept is *Data*, as we consider that risks are getting higher with the amount of collected data, and that is why we suggest scoring this concept according to the following rule:

If (individuals of *Sensitive_Data* is not null) then $RiskScore(Data) = 2$, if individuals of *Personal_Data* or its subclasses is not null) then $RiskScore(Data) = 1$, else $RiskScore(Data) = 0$.

4 Usage Scenario and Discussion

To demonstrate our approach, we analyzed the privacy policy of the August company that produces smart lock, doorbell cameras and other accessories [17]. Their smart lock allows implementing a variety of convenient but privacy risky functions as remotely lock and unlock the door, logging exit/entrance activity of smart lock owners as well as their guests, supports biometrical identification and voice assistant. We constructed an ontology for the privacy policy concerning August services and products [17] and calculated privacy risks based on the information provided within it.

We examined the following data usage aspects: *first-party data collection and usage scenario, third-party data collection and sharing, data security, data retention, privacy settings, data control and policy change*. We omitted from the explicit risk analysis international and special audience scenario as special audience is usually represented by citizens of EU and California protected by a set of regulations such GDPR [1], CCPA [18].

These regulations require specifying explicitly the purpose of data processing including collection and third party sharing, and our analysis showed that these concepts are considered in first data collection and third party-sharing. Moreover, it is interesting to understand privacy risks in general but not for a specific audience, however, in future we are planning to include this scenario and analyze the difference in privacy risks for different type of audience. The August products are not purposed for the use by minors under 16, therefore privacy risks for this specific audience are not calculated. It also should be noted that the usage scenario describing *privacy breach notification* was not detected in the text at all.

The given privacy policy is represented by a text structured as a sequence of paragraphs, some of them contain bulleted lists, there was also a table. Currently the process of mapping privacy policy to P2Onto concepts is done manually, however, to detect P2Onto concepts and data usage scenarios in text of the policy we used assumptions defined in Sect. 3.2 and treated paragraph without bulleted list or item of a list as one data usage scenario.

Figure 5 shows a part of constructed P2Onto ontology describing collection activity constructed under following assumptions.

We used Graffoo OWL editor [19] for prototyping and visualizing ontology before moving it to OWL/XML format. We also used its capabilities to highlight different usage scenario for different type of data. The rectangles on Fig. 5 correspond to P2Onto concepts, labeled arrows – to object properties, while small circles – to the individuals. The individuals that belong to one usage scenario, i.e. were detected in one paragraph or bullet list item are marked by one color.

From Fig. 5 it is clearly seen that in major cases the privacy policy text did not contain individuals of all P2Onto classes referring to one data usage, this resulted in appearance of *NotDef* individuals in many data usage scenarios. In some cases these concepts were described separately. For example, the purposes of data usage and storage were given in separate paragraph, but there was no clear specification what type of the data they refer.

We detected similar case in data retention usage practice (Fig. 6).

The product manufactures first provide general description of how long all collected data is retained for, the purpose and legal basis for data retention (white circles in Fig. 6),

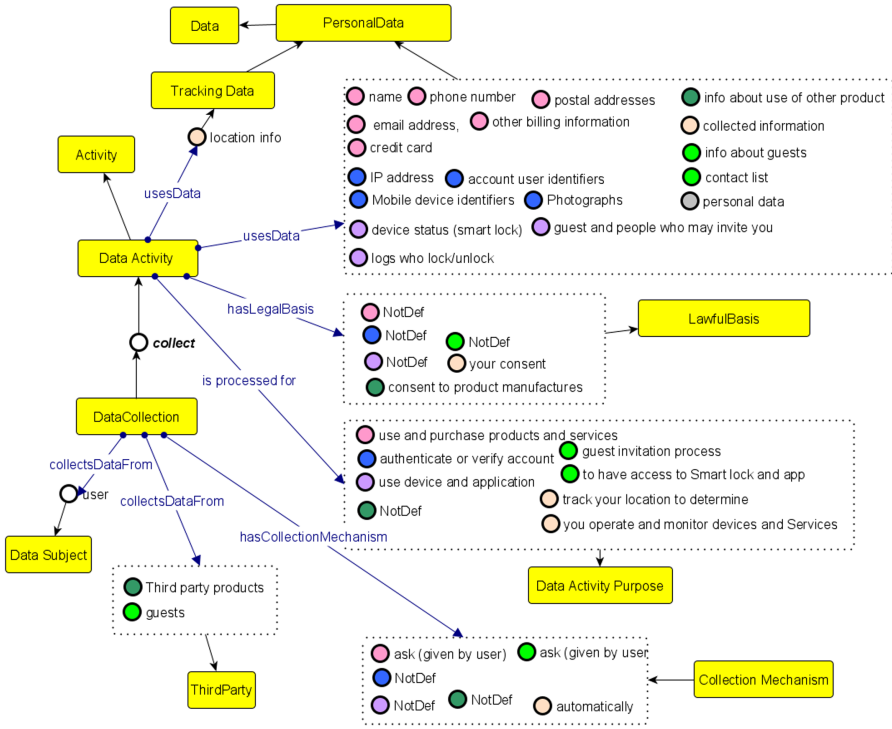


Fig. 5. Part of P2Onto ontology presenting first party collection data usage practice detected in August privacy policy (Color figure online)

and then specify some particular scenarios, for example, they inform that lock activity, including guest activity as well as account information is stored at least 90 days after account deletion (orange circles in the Fig. 6), however, they do not provide information how long the financial data of the smart lock user is stored (lilac circles in the Fig. 6).

Let us calculate privacy risks for first party collection and usage practice. It is described by five competence questions that are given in Table 2 alongside with corresponding P2Onto concepts. To calculate risk score associated with each competence question it is necessary to assign ranks to the individuals detected.

Table 3 contains suggested ranks for the individuals.

We assumed that collection information from the user’s guests may pose high privacy risks both to user and his/her guests. The purposes of data processing concerning personalization and understanding of user behavior are also considered as critical as they highly related to user behavior profiling, and at last we refer to *not defined* legal basis of data collection and processing as critical, as personal data processing has to have clearly defined basis for this activity what is stated in many legislative regulations.

For the assigned ranks of the individuals we obtained the following risk scores for each competence question or corresponding P2Onto concept and risk score for the given data practice (see Eq. 1):

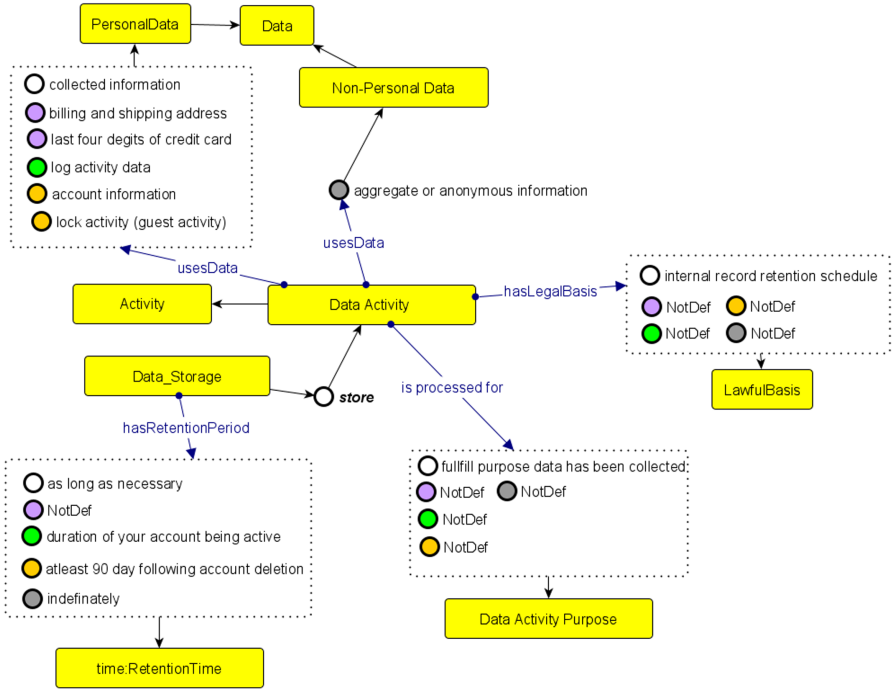


Fig. 6. Part of P2Onto ontology presenting data retention practice detected in August privacy policy (Color figure online)

- $RiskScore(Data) = 1;$
- $RiskScore(Third\ Party) = 2;$
- $RiskScore(Collection\ Mechanism) = 0;$
- $RiskScore(Data\ Activity\ Purpose) = 2;$
- $RiskScore(Legal\ Basis) = 2;$
- $RiskScore(First\ party\ collection\ and\ usage) = 7.$

The values of the risk thresholds for the given usage scenario are the following:

- $High_Threshold = 4/3 \cdot 5 = 6.67,$
- $Low_Threshold = 2/3 \cdot 5 = 3.33,$

Thus, The *RiskScore* equal to 7 corresponds to high privacy risks.

It should be noted that the procedure of assigning ranks to the individuals is a critical part in the risk assessment procedure. For example, changing rank of *not defined* legal basis to not critical results in *RiskScore* for *Legal_Basis* concept equal to 0, that in its turn results in medium privacy risks for this data practice (*RiskScore* = 5).

Table 3. Assigned ranks of P2Onto individuals for different concepts

P2Onto concepts	Critical	Generic	Others
Data	All individuals of Sensitive_Data class	All individuals of Personal_Data class	All individuals of Non_Personal Data
Third party	Guests	Third party products	
Collection mechanism			Automatically ask (given by user)
Data activity purpose	Personalized, understand your needs, interests	Use and purchase products and services; use device and application; provide, administer, improve app, services; provide service; product and communication; comply legal obligations; enforce agreements; provide further information and offers	Guest invitation process, to have access to Smart lock and app, authenticate or verify account, track your location to determine..., You operate and monitor devices and services conduct market research, guest invitation process, manage and administer our account, fulfill orders, respond to support requests; resolve disputes; protect, investigate, deter against fraudulent, illegal activity; administer promotional activity
Legal basis	NotDef		Your consent, consent to product manufactures, expressed consent, performance, our legitimate interest, legal obligation

Application of similar procedure to assess the rest of privacy policies allows obtaining following risk scores for them:

- $RiskScore(third\text{-}party\ data\ collection\ and\ sharing) = \text{Medium}$
- $RiskScore(data\ security) = \text{Medium}$
- $RiskScore(data\ retention) = \text{High}$
- $RiskScore(privacy\ settings) = \text{Medium}$

- $RiskScore(data\ control) = \text{Medium}$
- $RiskScore(policy\ change) = \text{Medium}$.

The overall risk score for the given policy is Medium, that it is expected privacy risks for this policy, as the device collects and stores a lot of personal information that relates not only to the end users but to their guests. However, the sharing process is described rather transparent, though the format of data sharing is not defined.

Interestingly that it is clearly stated only when data sharing is done in market research and other purposes. The retention data aspect received High risk score because it has indefinite period of retention, however this period is mentioned in the usage scenario of aggregated and anonymous data.

This made us to conclude that it is necessary to consider the type of the data (personal, sensitive or non-personal data) involved in each data scenario. The application of the ontology as a framework for constructing such rules allows these changes as all data scenarios are presented as linked ontology concepts. This ability of the ontology is also useful in explaining obtained results as it is clear how different types of personal data are collected, processed and shared, what tools and options to access, edit personal data or delete of them are available to end user, etc.

The authors consider that this ontology can serve as the basis for elaborating interactive graph-based visualization models targeted to explain the privacy risks to the end user in clear and readable manner.

5 Conclusions

The personal data protection is highly relevant task in the modern information systems due to their complexity and strong link with everyday life of people, on the one hand, and possible negative consequences of the personal data leakage, on the another hand. In some cases privacy polices are the only way for the end user to understand what types of personal data are processed by device or application, how they are processed and protected, what the goals of data collection and sharing are.

This paper proposed an approach for privacy risk assessment based on ontology constructed for a particular privacy policy. The risk assessment procedure uses rules that score privacy risks depending on the rank of ontology individuals detected in the text of privacy policy. The resulting scores can help end user to understand what privacy risks he/she accepts when accept privacy policy.

In the paper the authors demonstrated the proposed approach for assessing privacy policy of the smart lock that allows remote control. The usage scenario showed that proposed ontology is able to present main data usage aspects in clear and readable manner, it also allows explain the calculated risk score.

However, it also revealed that setting ranks for individuals is a critical aspect that requires additional research. Another important direction of the future research is related to the automation of ontology concepts detection in the policy text.

References

1. General Data Protection Regulation (GDPR). <https://gdpr-info.eu/>
2. Ashley, P., Hada, S., Karjoth, G., Schunter, M.: E-p 3p privacy policies and privacy authorization. In: Proceedings of the ACM workshop on Privacy in the Electronic Society (WPES 2002), Washington, DC, USA (2002)
3. Karjoth, G., Schunter, M.: Privacy policy model for enterprises. In: Proceedings of the 15th IEEE Computer Security Foundations Workshop, Cape Breton, Nova Scotia, Canada (2002)
4. Ardagna, C.A., De Capitani di Vimercati, S., Samarati, P.: Enhancing user privacy through data handling policies. In: Damiani, E., Liu, P. (eds.) DBSec 2006. LNCS, vol. 4127, pp. 224–236. Springer, Heidelberg (2006). https://doi.org/10.1007/11805588_16
5. Pardo, R., Le Métayer, D.: Analysis of privacy policies to enhance informed consent. In: Foley, Simon N. (ed.) DBSec 2019. LNCS, vol. 11559, pp. 177–198. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22479-0_10
6. Gerl, A., Bennani, N., Kosch, H., Brunie, L.: LPL, towards a GDPR-compliant privacy language: formal definition and usage. *Trans. Large-Scale Data- Knowl.-Centered Syst.* **37**, 41–80 (2018)
7. De, S.J., Le Metayer, D.: Privacy risk analysis to enable informed privacy settings. In: 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), London, pp. 95–102 (2018)
8. Zimmeck, S., et al.: MAPS: scaling privacy compliance analysis to a million apps. In: Proceedings on Privacy Enhancing Technologies, vol. 66 (2019). https://ir.lawnet.fordham.edu/faculty_scholarship/1040
9. Kumar V.B., et al.: Finding a choice in a haystack: automatic extraction of opt-out statements from privacy policy text. In: Proceedings of the Web Conference 2020 (WWW 2020), p. 1943–1954. Association for Computing Machinery, New York (2020)
10. Oltramari, A., et al.: PrivOnto: a semantic framework for the analysis of privacy policies. *Semant. Web* **9**(2), 185–203 (2018)
11. Children’s Online Privacy Protection Rule (“COPPA”). <https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule>. Accessed 05 July 2020
12. Health Information Privacy. <https://www.hhs.gov/hipaa/index.html>. Accessed 05 July 2020
13. Pandit, H.J., O’Sullivan D., Lewis, D.: An ontology design pattern for describing personal data in privacy policies. In: WOP@ISWC (2018)
14. IoT Security Compliance Framework. <https://www.iotsecurityfoundation.org/best-practice-guidelines/>. Accessed 05 July 2020
15. GSMA IoT Security Guidelines and Assessment. <http://gsma.com/iot/iot-security/iot-security-guidelines/>. Accessed 05 July 2020
16. PROV_O: The PROV Ontology. <https://www.w3.org/TR/prov-o/#Agent>. Accessed 05 July 2020
17. August Device and Service Privacy Policy. <https://august.com/pages/privacy-policy#product>. Accessed 05 July 2020
18. California Consumer Privacy Act 2018. <https://oag.ca.gov/privacy/ccpa>. Accessed 05 July 2020
19. Graffoo OWL Editor. <https://essepuntato.it/graffoo/>. Accessed 05 July 2020