Towards a User Centric Personal Data Protection Framework

Hind Benfenatki
University of Lyon, CNRS,
INSA-Lyon
LIRIS UMR 5205
Lyon, France
hind.benfenatki@insa-lyon.fr

Frédérique Biennier
University of Lyon, CNRS,
INSA-Lyon
LIRIS UMR 5205
Lyon, France
frederique.biennier@insa-lyon.fr

Marco Winckler
Université Nice Sophia
Antipolis
Nice, France
winckler@unice.fr

Laurent Goncalves
Softeam
Business Unit e-Citiz
Toulouse, France
lgoncalves@e-citiz.com

Olivier NICOLAS
Softeam
Business Unit e-Citiz
Toulouse, France
onicolas@e-citiz.com

Zohra Saoud
Université Claude Bernard
Lyon 1
LIRIS UMR 5205 CNRS
Lyon, France
zohra.saoud@liris.cnrs.fr

ABSTRACT

Nowadays, data has become marketable, and users often underestimate the impact of not protecting their data. Even more than that, some users agree to sale their private data. In this context, data protection and user sensitization have to be at the heart of our concerns. This work is part of Personal Information Controller Service (PICS) project. In this paper, we propose a user-centric data protection service to (1) allow users to identify their protection requirements thanks to risks evaluation support tools, (2) provide a "privacy evaluation" of SaaS suppliers based on their ToS and (3) allow users control the access authorizations they grant to SaaS providers. To this end, we define a semi-structured Personal Information System organization used to organize access rights based on a General Data Protection Regulation-compliant ontology.

Author Keywords

Data Privacy; Personal Information Controller Service; Personal Data Worth; Data Market; Privacy-by-design

INTRODUCTION

General Data Protection Regulation (GDPR) strengthens and unify data protection for individuals in the European Union [1]. The burden of proof is now on the provider and no longer on the person concerned. This can lead to two situations. First, the number of complaints can quickly increase given that the rest of the procedure has to be done by service providers. This will make the juridical and IT divisions solicited more regularly leading to the creation of job profiles which mission is to treat this requirement. This implies the reorganization of organization's functioning. Second, in order to protect themselves, providers may describe all the actions to be done on user's personal data in Terms of Service (ToS) knowing that users do not read it. To overcome these shortcomings, one can intervene before using the targeted services by evaluating automatically all the charters described in ToS in order to determine if it meets user's protection requirements. To do that, ToS and user's Requirements of Protection (RoP) have to be described in a structured manner and to include protection concepts described in GDPR such as the data portability, the right of oblivion, the user consent and so on. To address this issue, we propose a GDPR-compliant ontology to describe provider's Quality of Protection (QoP) and user's requirements of protection. Set of questions have been designed to evaluate users' privacy awareness, risks aversions and their current practices. User's requirements are gathered in a dynamic way in order to make the user aware of lax requirements.

The rest of this paper is organized as follows. Section 2 describes the state of the art regarding personal data protection in the actual context. Section 3 describes the multi-layer personal information system's organization. Section 4 describes GDPR-compliant ontology for describing quality and requirements of protection. Section 5 describes risks management. Section 6 draws final conclusions.

CONTEXT AND RELATED WORK

Personal data protection requirements and motivation

Personal data is defined by GDPR [2] as *any information relating to an identified or identifiable natural person*. It regroups financial data, administrative data, identity data (e.g., name, date of birth), medical data, biometric elements (e.g., fingerprint), connection data (e.g., IP address), localization data (e.g. GPS tracking), activity data (e.g. cookies) and so on.

According to way it is obtained, personal data can be classified as explicit, collected, and generated data. The first

- License: The author(s) retain copyright, but ACM receives an exclusive publication license.

one represents the information given explicitly by the user (e.g., while filling out a form).

The second one represents the data induced by user's activity (for instance, connection activity, or even GPS coordinates). In addition to the use of different personal data, the identification of an individual may also arise from his/her interaction with different services. Then we have an identification from the traces of activity. While these identifications may seem less intrusive because they are not directly related to sensitive information, they nonetheless represent a major risk of privacy violation insofar as cross-linking of data sources can make it possible to link these identities derived from trace data, to a legal identity without the user knowing that he/she is identified.

The later one represents data resulted from information construction or data fusion. In fact, big data algorithms impose to cross many data sources, thus generating information not explicitly given by the user. Furthermore, some providers analyze user's personal data in order to construct an information which is not given explicitly. For example, receiving a travel ticket via Google mail service leads to email analysis so that travel information can be sent and displayed on the user's Android phone.

Personal data collection and analysis has to be taken into account while choosing a service. Usually, service' providers describe the actions done with personal data in the charters of ToS. However, on the one hand, ToS are too long to be read [3], and users often agree without analyzing the content of ToS. On the other hand, providers do not describe on ToS all the actions they do with personal data.

Identifying risks associated to personal information

According to a "security" point of view, users can use different services to ensure a "backup" of personal data. These services are sometimes inseparable from products (e.g., mandatory registration via iTunes / iCloud for Apple, authentication with a Gmail account for Google, etc.). Most of these service providers operate on a global scale and use providers to synchronize user data. The outsourcing contracts contain the policies for personal data protection, but when user data comes from diverse sources, there is always a risk of violation of the original suppliers' Terms of Service (ToS).

Physical mobility of might also increases the risk of personal data disclosure because continuity of service on the move requires users to communicate their identifiers to many services. On December 20th 2016, a line statement was included in ESTA forms completed by foreign tourists not having a visa before his/her arrival on US soil so that user credentials could be communicated on social networks (such as Facebook, Twitter, YouTube, Instagram, Google+., Vine ...). Coupled to the Patriot Act that enforces Internet service providers to transmit the information in their possession, this addition presents a significant risk of privacy violation.

Furthermore, many applications might keep data collected against users' will. We can take the example of the fight of the EU against Google's search engine for the right to be forgotten. Since 2014, European laws allow Europeans to request the deletion of personal data available on Internet concerning them and they consider obsolete or harmful to their image. Google, yet American society, had to comply with this decision and made available to European Internet users a form allowing them to make a request for deletion. At the beginning of the year 2017 (03/01/2017) 668,653 requests, for a total of 1,840,364 URLs, have been made since the launch of the procedure (29/05/2014). France is also the first country in terms of the number of applications with 217,852 applications, far ahead from the British, 2nd with 101,508 applications. However, all these requests are not accepted by Google, only 43.2% are (48.8% for France, 2nd as a percentage of acceptance behind the Czech Republic where 50.8% of applications are successful). The EU has managed to bend the big American firm ... but in fact, this result must be relativized since in reality the deleted URLs are only on European versions of the search engine (google.fr, google.uk ...) and not on others (google.com, google.us ...).

In addition to that, some service providers might track personal data without users being aware of it. Google recently admitted that they collected GPS data to enhance the provided services even when users turned off their GPS. Moreover, users were not informed that GPS data was being collected by Google.

Even when informed about the Terms of Services, users become major source of vulnerability when they provide "for free" sensitive personal information to SaaS providers without taken into account the risk of losing the control on each piece of their personal information. This anarchic dissemination of pieces of personal information makes harder the discovery of privacy failures for both users and SaaS providers who may be charged by users for "unfair usage of their personal information".

While security engineering methods have been designed for traditional information system, there's no mean to protect the unstructured personal information system nor to define protection requirements.

MULTI-LAYER PERSONAL INFORMATION SYSTEM'S ORGANISATION

As we shall see, personal data protection is a complex issue and the prevention of many threats requires some kind of user involvement. To overcome these shortcomings, we propose a user-centric model that provides:

(1) An open and up-to-date personal data description framework, including personal information description and their relationships with the different SaaS providers who accessed them. This element of the framework is aimed at inform users about SaaS which requires their personal data.

(2) A risk evaluation method based on user's awareness, information sensitivity and SaaS supplier protection policy. This element is aimed at informing users about the risks of granting SaaS personal data.
(3) A global dashboard synthesizing the targeted protection level compared to the (may be inconsistent) authorization granted to the SaaS providers. This element is aimed at recalling users about the current status of their personal data.

This model will rely on an automatic evaluation of services' quality of protection in order to determine if they match users' requirements. To this end, a GDPR-compliant structured definition of charters must be defined to identify potential risks from services' ToS.

Our architecture of the personal IS is organized in 3 layers:

- Persistence: these are storage services (digital vault ...) from which data can be exchanged. To extract a particular content, one must have access key defined by the provider (URI ...),
- Personal object: this layer includes the semantic description of the data. This is the heart of the PICS system insofar as PICS does not store personal data but allows to organize the storage and exchange of the user's personal data. The personal object is designated as a group in the modeling architecture. This object can be considered as an aggregated data which allows to describe it by a set of descriptors. In this way, authorization management can be trivialized by considering only the data instead of differentiating the data of the groups. The relation of data belonging to a group makes it possible to have different levels of description. For example, a group describing a postal address will be associated with the class "address" of the pattern "addresses and contact points". It may contain more specific groups or elementary data allowing this address to be broken down according to the desired level of structuring. The group is not directly associated with a content, it is the data associated with the group that allows the association with the content and therefore with the persistence model (the storage service (s)). It is possible to choose a main storage and several secondary storages.
- Application - usage: this layer will be used to store information about SaaS services accessing personal information. The current model links the SaaS service to a provider. The access request relates to a given data.

This structured personal information system model allows defining the protection of personal data as in classical Information System. A security policy is associated with each data and allows to define security requirements according to the classical security services (i.e. confidentiality, availability, integrity and non-repudiation). The privacy requirements integrate:

- The sensitivity level for a data. This makes it possible to establish a classification into simple categories of data that can facilitate the creation of access policies to the data,
- Circles of trust on service providers based on the sensitivity level of the data,
- The profile of the user to be able to propose adapted security policies, through a questionnaire to understand not only the usual behavior of the user in the digital world and its "security awareness level", but also its digital risk aversion.

The need for integrity requires that the data (stored or contained in messages exchanged with a service) are those expected and have not been tampered voluntarily or inadvertently.

The need for availability is to ensure that the data (or the service to access it) is available and functional at any time. This last need may also be related to quality of service guarantees.

The need for non-repudiation is to ensure that no user or service provider can deny or challenge the actions or operations performed. This need requires keeping track of the activity and providing certified exchanges for users to have evidence.

The need for confidentiality consists in ensuring that the data is only accessible to those who have access authorization. This need concerns both the protection of the data contained in the messages exchanged between the various actors and the preservation of the security of the data hosted by a service provider. This need uses several sub-criteria:

- Authentication consists in ensuring that the user is the one who requests access to the data.
- Authorization consists of ensuring that an authenticated user has the necessary rights for and / or is authorized to access the data.
- Encryption consists in making unintelligible a data to any person having no right of access, as only authorized users have the decryption key.

This privacy policy is designed to fit the GDPR requirements. To this end a GDPR-compliant ontology is needed.

AN ONTOLOGY FOR DESCRIBING QUALITY AND REQUIREMENTS OF PROTECTION

In order to assess if the quality of protection provided by SaaS meets the users' requirements of protection regarding actions on data, we must at first make sure that data privacy concepts are non-ambiguous and can be seamlessly applied to all applications. For that, we propose a GDPR-compliant ontology describing concepts related to data privacy usually described in terms of service [4] and in GDPR. Thereby, it is possible to confront automatically user's requirements of protection to the encountered risks while analyzing services' quality of protection.

Our ontology is composed of three main concepts as illustrated in Fig. 1: (1) the service class which is a Linked Unified Service Description Language (USDL) class [5], [6], (2) the ProtectionConcept class gathering actions on data, governance parameters, security concerns, and user's training (see Fig. 2a), (3) the ProtectionAttribute class which includes attributes related to protection concepts such as systematic user notification while transferring one of his/her information (Fig. 2b).

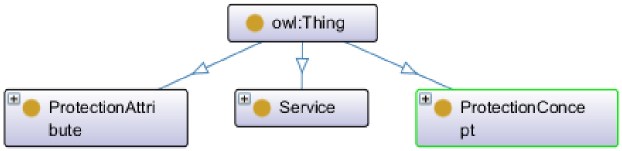

Figure 1. GDPR-compliant ontology main classes

For each kind of data, action on data in the GDPR-compliant ontology describes the actions the provider can do on it. Personal data are classified in section 2. If an action on data is described for a general data type (for instance, financial data), then specific data types (for instance the International Bank Account Number) inherits de facto the description of the action. The opposite is not true.

Our goal is to put the protection of personal data according to a classic vision of the security of traditional IS. For this, the modeling framework attaches a security policy description to the different personal data. Each piece of data is characterized by a level of sensitivity and a protection.

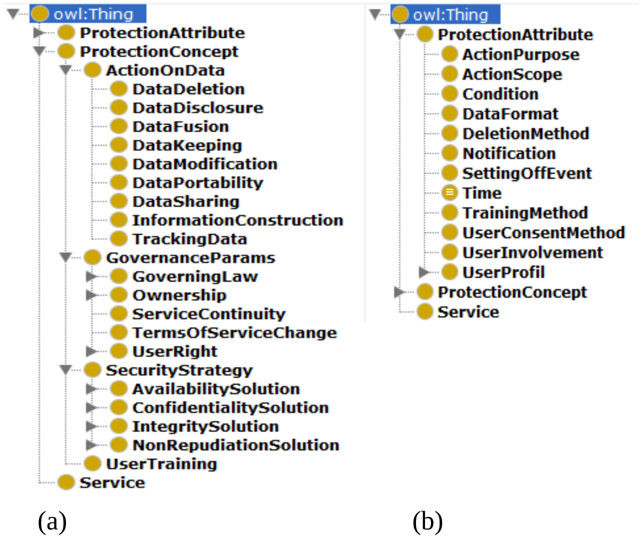

Figure 2. (a) GDPR-compliant ontology protection concept, (b) GDPR-compliant ontology protection attributes

More precisely, actions on data include topics and concepts related to user's data usage such as:

- Data sharing: when service providers share users' personal information with other stakeholders, GDPR states that users must at least consent to such transfers, including even the type of shared data and the purpose of such exchange (a business transfer, a legal obligation, etc). The user can also be notified after the sharing operation occurs. Depending on legal context, the user consent may be avoided (homeland security, particular agreements). In our work we consider four attributes to qualify these transfers based on TeleManagement Forum [7] service evaluation: user's consent method, user's notification, sharing scope, and sharing purpose.

- Tracking data occurs when user's personal data are used for other purposes than originally intended. Tracking data action is evaluated in our work regarding four attributes which are: user's consent method, user's notification, tracking scope, and tracking type (e.g., GPS information, research keywords).

- Information construction occurs when user's raw data are analyzed to deduce information. Such operation is qualified using user's consent method, user's notification, and action's scope (which data are analyzed).

- Data modification allows the user to modify its personal data. Three attributes are associated to this operation which are: modification scope (i.e., which data will be modified), user's notification, and action's trigger (i.e., the required action to modify personal data such as sending a request by email, directly by clicking in service's website, etc).

- Data fusion occurs when the current service merges the data given by the user with other collected data. This leads to provide more information to the service provider than the ones that are explicitly given. For example, a Facebook based connection on a professional social media can lead the provider to collect your Facebook friends. This may be intrusive since after that the service knows more information than what was explicitly given. Three attributes are associated to this concept which are: fusion scope (i.e., which data will be collected and merged), fusion purpose, and user's notification.

- Data keeping describes how much time the provider keeps user's data.

- Data deletion is related to the right to be forgotten. Related to data keeping period specification, data deletion should occur when users' data are no longer used. For example, when you apply for a job, your application is processed in a limited time. After the selection occurs your application should be removed. Some services keep your resume for one year. The deletion can also be triggered by the user before the retention delay is reached. We describe four attributes for this concept which are: retention period, deletion method (e.g., automatic, manual), triggering event, and notification.

- Data portability allows users to get their data back in an interoperable format when they decide to move from one service provider to another one. For example, Twitter and Facebook data can be retrieved as an XML file archive. Data portability can be triggered by a user's action, such as a request by email, or can be done automatically when leaving the service. Three attributes are associated to this concept which are: data format, user's notification, and action's trigger.

Terms of service and governance concerns are associated to a set of rules, conventions, and contracts to improve stakeholders' coordination. This includes:

- Jurisdiction and governing laws.
- User rights: allows to describe the leaving right and the rights in which the user waive. For example, by agreeing to Instagram's terms, you agree that you can only bring any claim against them on your own behalf. You waive your right to being part of a class action, or participating in claims brought in any representative capacity [8].
- Ownership describes who owns the introduced data and the generated data via a copyright.
- User involvement in changing terms: describes if changes are proposed as a request for feedback, or imposed unilaterally as a take-it-or-leave-it deal.

Security Policy is related to data availability, confidentiality, integrity and non-repudiation requirements and associated solutions. The protection of the user privacy consists in preserving the anonymity of the user by ensuring that he cannot be identified, whether via his/her identity, his/her location or any other trace of activity. The protection of privacy also requires securing personal data whether it is the messages exchanged or the processing of these data. This last point requires the ability to manage the user's consent, particularly with regard to the disclosure of personal data to a third party. However, the guarantee of anonymity is problematic in a context of statistical processing (authorized by the latest European directive) because big data algorithms impose to cross many data sources and the anonymization of data can no longer be guaranteed in this "open" context. Finally, the legal provisions related to the protection of personal data may vary depending on the location of data storage.

Stakeholder training describes the formation level of providers' stakeholders to protect users' data.

RISKS MANAGEMENT

As said previously, the privacy policy integrates data sensitivity, provider's trust level, and user's profile. This last point is related to both uses security awareness and to security aversion. In order to define an adapted user-centric security and privacy strategy, we propose a privacy requirements capture method based on several steps: user profile identification, user's requirements of protection gathering, and the evaluation of user's protection level regarding currently used services and targeted ones.

User profile identification

User security awareness and risk aversion are captured thanks to a set of questions/answers. These questions allow the system to identify how the user apprehends his/her personal data and what are his/her practices / criteria for sharing these information. We focus on different axes:

- Mastery of the concept of identity and digital identity: "how to define how a user can be recognized and by whom"? This goes through the identification of which kind of data the user publishes on the Internet, the habits regarding the authentication (e.g., the use of biometric authentication, password changes), and traces (e.g., cookies),
- Mastery of what is personal data: this is to allow him/her to understand the "what to protect"? This goes through the identification of the actions the user does in online services implying the provision of his/her personal data,
- Mastery of the legal framework: this is to define the understanding of the charters, the rights and uses they include, licenses, and data property. It is the "to do what with my data"?
- Technological mastery involves addressing the level of technical knowledge and technical practices of the user vis-à-vis technological tools to access services including security practices.

Table 1 illustrates a question related to identity data management. The possible answers are classified from strong to weak constraint.

Table 1. Question related to identity data management and its response classification

Question	Strong constraint	Neutral constraint	Weak constraint
Do you agree to provide your first and last names?	Only if I am obliged to do so and that guarantees are given to me	Provided it seems necessary to me	No problem

The answers obtained will allow us to evaluate the level of sensitivity of the user to the protection of personal data, and so to define the user's profile.

User's requirements of protection

Requirements of protection's description follows the GDPR-compliant ontology described in Section 4. A questionnaire is used to allow the user to describe its requirements of protection. To this end, we propose also a set of questionnaire which browses all the ontology's concepts. User's requirements of protection are gathered in a dynamic way, i.e., according to user's requirements and user's profile. For each requirement, if a user having a weak IT-hygiene level, provides a lax answer (weak constraint), the user is notified and an explanation is provided to raise his/her awareness in order to make him/her change his/her answer.

Evaluation of current protection level

In order to evaluate the current protection level of the user regarding the already used services, we aggregate all quality of protection of used services QoP(Si) (i=0, i<number of used services) using the negative aggregation. QoP_A represents the aggregated QoP. Negative aggregation consists of describing the worst situation regarding quality of protection for the same protection concept. As an example, let us suppose that QoP(S1) states that data transfer is not allowed without user's consent, and QoP(S2) states that user's data are transferred without any user's consent. In this situation the aggregated value states that no consent is required before transferring user's data.

QoP_A is then compared to user's RoP in order to confront the user to its actual protection. The process is also done for each quality of protection of a given service QoP(Si) with user's requirements in order to determine the questioned services.

After that, first, the user is confronted with the worst protection level he agreed to, represented by QoP_A. Then, the services which do not match user's requirements of protection are highlighted and non-respected requirements are brought out, so that the user can decide knowingly to leave the service or to continue to use it.

User's requirements vs provider's quality of protection

While a user is targeting a service, risk analysis is done in two stages: risk regarding user's expectations, and risk regarding the ideal situation. In the former, service's quality of protection is compared to user's requirements. Matching mechanisms are used in order to highlight and brought out each quality of protection which does not comply with the corresponding requirement. Here, noncompliance means that the quality of protection is more permissive than required. This is deduced by comparing for each considered protection concept, the range of its attributes in the corresponding reference table. As a reminder, the lower the range is, the better the protection is.

In the later, the ideal situation corresponds to the better attributes' values in the corresponding reference tables, for each considered protection concept. Even if the quality of protection matches user's requirement, if it does not match the ideal situation, the user is notified and sensitized so that the user can decide knowingly to leave the service or to continue to use it.

This analysis will provide a service score compared to the current protection score and the requirements of protection. In this way, the user can decide if the access to the data can be granted or not.

CONCLUSION

This work is part of a project that investigates strategies for granting users control and protection of personal data exchange. Our approach integrates tool support for managing personal data, setting permissions, detecting data transfer SaaS, and prompting users about vulnerabilities. In this paper, we presented the impact of the application of GDPR both from provider and user perspectives and described how to deal with the induced changes. As we shall see, the protection of personal data not only depends on use of services but also requires users understanding and control of their personal data. The ambiguity and complexity of concepts that are used to define data protection is an issue that we approached in this work by proposing an GDPR-compliant Ontology. Such as ontology can be used to describe user's requirements and provider's quality of protection. In fact, the use of structured description of ToS allows to analyze automatically the matching between requirements and quality of protection. We also integrate the analysis of user's profile in order to guide and to sensitize the user while targeting a service and so personalize the user orientation. As future work, we are planning to perform user testing of some tools that we are developing as proof of the concept. We also want to discuss the use of the GDPR Ontology in tools aimed at helping users to protect their data.

ACKNOWLEDGMENTS

This work is partly supported by the Personal Information Controler Service (PICS) project, co-sponsored by the French Secrétariat Général pour l'Investissement, the French Direction Générale des Entreprises and Bpifrance under the "Investissement d'avenir - Protection des données personnelles" grant.

REFERENCES

1. "Règlement européen sur la protection des données : ce qui change pour les professionnels," 15 juin 2016. [Online]. Available: https://www.cnil.fr/fr/reglement-europeen-sur-la-protection-des-donnees-ce-qui-change-pour-les-professionnels.
2. "GDPR, Art.4" [Online]. Available: http://www.privacy-regulation.eu/en/article-4-definitions-GDPR.htm.
3. "Terms of Service, Didn't Read," [Online]. Available: https://tosdr.org.
4. "Topics - Terms of Service, Didn't Read," [Online]. Available: https://tosdr.org/topics.html#topics.
5. J. a. C.Pedrinaci, "Evolution and overview of Linked USDL," in *International Conference on Exploring Services Science IESS*, 2015.
6. "Linked USDL modules," [Online]. Available: https://github.com/linked-usdl.
7. "TeleManagement Forum," [Online]. Available: https://www.tmforum.org

1. [Online]. Available: https://tosdr.org/topics.html#waiver-KoJc2tTrdFQ