PREPRINT
Final article published in: Society, Privacy and the Semantic Web - Policy and Technology (PrivOn 2017), co-located with ISWC 2017
proceedings: http://ceur-ws.org/Vol-1951/PrivOn2017_paper_6.pdf

Modelling provenance for GDPR compliance using linked open data vocabularies

Harshvardhan J. Pandit and Dave Lewis

ADAPT Centre, Trinity College Dublin, Dublin, Ireland

Abstract. The upcoming General Data Protection Regulation (GDPR) requires justification of data activities to acquire, use, share, and store data using consent obtained from the user. Failure to comply may result in significant heavy fines which incentivises creation and maintenance of records for all activities involving consent and data. Compliance documentation therefore requires provenance information outlining consent and data lifecycles to demonstrate correct usage of data in accordance with the related consent provided and updated by the user. In this paper, we present *GDPRov*, a linked data ontology for expressing provenance of consent and data lifecycles with a view towards documenting compliance. *GDPRov* is an *OWL2* ontology that extends *PROV-O* and *P-Plan* to model the provenance, and uses *SPARQL* to express compliance related queries.

Keywords: privacy, ontology, GDPR, compliance, consent, provenance

1 Introduction

The General Data Protection Regulation (GDPR) [7] introduces important changes to the way data is obtained and processed, and is set to come into effect next year from 25$^{\text{th}}$ May 2018. An important change from previous data regulations is in the requirements for consent regarding providing information to the data subject about the data collected and any intended use including storage and third party sharing. The approach of using a lengthy and legal-speak terms and conditions with a check-box at the bottom to accept terms as given consent is no longer a valid mechanism under the GDPR. Instead, it is necessary to demonstrate that consent was obtained explicitly (for sensitive data) and in an unambiguous manner while clearly stating the data to be collected, the purpose of its intended use, and any third parties it is shared with along with the purpose of sharing.

Compliance with the GDPR is important as prospective fines are set at up to 20 million euros or 4% of the organisation's global turnover, whichever is higher. Proving or demonstrating compliance requires maintaining provenance traces that can demonstrate how the data was collected, used, stored, and shared by various activities along with justification in the form of consent obtained. Similarly, the lifecycle of consent must also be recorded to provide provenance

metadata about how it was obtained and whether it was used correctly to process data in agreement with the conditions outlined in the acquired consent.

Any approach towards documenting such activities should be capable of recording and querying provenance traces with regards to GDPR compliance, while being able to model the consent and data lifecycles at an arbitrary level of granularity. Such a vocabulary would have to be explicitly tailored to express the terminology of the GDPR, and would have to be based on open standards that can be easily adopted and extended by the community. While there is no substitute for astute legal documentation, it is possible to ease the task of representing it using linked data vocabularies that can be queried using a language such as *SPARQL* that provides a standardised way to query for data. Such a vocabulary would be distinct from a permissions or a rights-management system in that it only caters to the representation of provenance traces over consent & data lifecyles. It can, however, form the necessary base for expressing constraints over the provenance in terms of compliance as well as some form of access control.

In this paper, we discuss how semantic web vocabularies can be used to express the provenance information required for managing GDPR compliance and introduce *GDPRov* - an OWL2 ontology for expressing provenance traces of consent and data lifecycles. We also discuss how compliance related queries can be expressed using *SPARQL* over consent and data lifecycles declared using *GDPRov*.

The contributions of this paper can be summarised as:

1. Identifying provenance information related to consent and personal data required for compliance documentation
2. *GDPRov* - a linked open data ontology to represent provenance of consent and data lifecycle workflows for GDPR
3. Using *SPARQL* to formulate compliance related queries over provenance expressed using *GDPRov*

The rest of the paper is structured as follows: Section 2 discusses the provenance requirements for formulating GDPR compliance queries. Section 3 presents the *GDPRov* ontology, and Section 4 discusses the use of *SPARQL* to formulate compliance related queries. Section 5 discusses the related work while Section 6 concludes the paper with a discussion about future work.

2 Provenance information for GDPR compliance queries

2.1 Consent

GDPR heavily emphasises a consent-based mechanism for all activities involving personal data. Consent is the permission or agreement specified by the user for actions involving their data, and under GDPR is considered to be valid only when freely given, is specific to the request, is unambiguous, and informs the user regarding the nature and purpose of processing and the entities involved. It is obligatory for records to be maintained that outline how the consent was

acquired and the data activities permitted under it. In case of change in consent, the previous consent must be archived or maintained as a record of retroactive justification for the acquisition and usage of data under it. As GDPR has specific clauses regarding modification or withdrawal of given consent, it is necessary to demonstrate that activities using consent do not infringe on these obligations by only using an up-to-date consent.

It is difficult to determine whether consent was correctly obtained due to clauses such as 'informed' and 'unambiguous' that cannot be verified using only the provenance metadata. Therefore, it becomes necessary to also record the mechanisms which were used to obtain consent along with all the entities which influenced the decision. These could include for example, a HTML form shown to the user to obtain consent along with terms and conditions specified at the time. Using provenance metadata, it can be determined which version of the form or terms and conditions was used for obtaining consent from the user. This can then be expressed as a log of records used as objective documentation for decisions on whether the consent was rightly obtained for activities in the past.

2.2 Third party sharing

GDPR makes it mandatory when acquiring explicit consent from the user to explicitly specify previously ambiguous references to 'third parties' - entities other than the organisation who are involved through sharing of data. It also requires specifying the nature of the data being shared and its purpose. Under GDPR, a *Data Processor* is defined as an entity that does not exercise responsibility over the data it holds. By contrast, a *Data Controller* must require consent to justify its activities involving personal data. There are specific rules regarding sharing of data between Data Controllers that mandate that each Data Controller must obtain its consent independently and directly from the user (or an agent authorised to act on their behalf) if it needs to use data outside the agreement provided by the given consent. Provenance actions that 'share' data or any activity that specifically sends data 'outside' the boundaries of the organisation must be specified as an identifiable entity along with the role it plays in relation to the data.

2.3 Data Collection & Usage

According to GDPR, all data collection must be justified using user consent obtained against the specified usage of data by activities. This means data obtained based on consent for a particular activity cannot be used in another activity without obtaining explicit consent that permits such usage. To demonstrate compliance towards this aspect requires consent to be obtained using the provenance metadata for activities. This can be done by formulating consent as an agreement over data involved, activities that will use that data, and terms under which data is given, stored, and shared. Provenance metadata tracking the origin, use, and sharing of data can be helpful in checking whether it satisfies the conditions of

the compliance agreement by comparing the obtained consent against its usage in the recorded provenance of activities referenced by the consent itself.

2.4 Data Storage

Consent must be obtained with the intended duration of storage for data. Indefinite storage of personal data is not allowed, and the controller must periodically renew the consent to continue storing data. In cases where data is archived, transformed, or combined with other data, the controller must mention this in the mechanism that obtains consent from the user. Data lifecycles must clearly be able to demonstrate how data was obtained or generated, its usage, including any transformation, and subsequent storage. The provenance metadata for such lifecycles also includes activities such as anonymisation and archival which can be considered as specific variations of transformation and storage respectively.

2.5 Anonymisation of Personal Data

If the user data is pseudo-anonymised, GDPR permits certain freedoms regarding its usage depending upon the degree and control of de-anonymisation possible by the data-holding organisation. If data is completely anonymised, which means that it cannot be de-anonymised, it can be used in any activity regardless of the consent under which it was obtained. But it is important to note that this is 'complete' anonymisation - that is, there is no chance of linking it back to the user even with additional data, and that there is no discernible means for de-anonymisation. GDPR also specifically mentions the scenario where data is pseudo-anonymised and the organisation does not have sufficient additional data required to de-anonymise it. In such cases, it is permitted to treat the data as effectively anonymised for use within internal activities i.e. such data may not be shared with third parties without explicit consent. In case of data deemed to be private and sensitive, it is considered good practice to store it in a pseudo-anonymised form as measures against unwarranted and unauthorised access.

Provenance metadata for data that goes through anonymisation must also contain 'degree of anonymisation' - an arbitrary property that states the possibility for de-anonymisation. This will allow introspection over whether data was effectively anonymised before sharing, or in the case of a data breach, allow identifying the form of data accessed. We express the degree of anonymisation based on [15] with four levels ranging from completely de-anonymised (or not-anonymised) to pseudo-anonymised that can be de-anonymised by the organisation, pseudo-anonymised that cannot be de-anonymised by the organisation, and completely anonymous.

2.6 Additional rights

Under GDPR, data subjects have certain rights for withdrawing consent, rectification of data, and requesting a copy of their data. These rights can be exercised

at any time, are mandatory for the organisations to follow, and need not be part of the functionality specified to the user during consent. Therefore, processes that handle or correspond to these rights need to be documented separately from other services provided to the user. The provenance metadata for such processes should be able to describe a detailed plan of execution of what takes place whenever the user chooses to exercise a particular right, as well as be able to demonstrate that obligations for providing that right were followed. For example, handling the right to provide the user with a copy of their personal data requires that the copy must not be in a proprietary format and must be portable. Therefore the provenance describing this process must also be able to state the format in which the data was provided to determine if it followed the obligations mentioned under the GDPR. This information must then be recorded as a provenance record of having followed that right, which can be retrospectively checked or demonstrated as proof for compliance.

3 *GDPRov*

GDPRov (pronounced as GDPR-prov) is an OWL2[1] ontology for describing the provenance of data and consent lifecycles using GDPR terminology. It extends the existing linked open data provenance ontologies - PROV ontology[2] (*PROV-O*) and Ontology for Provenance and Plans[3] (*P-Plan*). *PROV-O* is used to represent provenance information and is a W3C recommendation. GDPRov uses these provenance ontologies to express a data-flow model that can trace how consent and data are used by extending the appropriate vocabulary with GDPR-related terms. The following subsection provides a brief description of *PROV-O* and *P-Plan* and how their core models map into *GDPRov*. The later subsections describe the core model of how *GDPRov* models concepts discussed in Section 2 using provenance ontologies. The final subsection describes how these concepts are instantiated as executions for representing real-world use cases. The OWL2 vocabulary for *GDPRov* ontology is available online[4].

3.1 *PROV-O* and *P-Plan*

Provenance is information about entities, activities, and people (or software) involved in producing data or a component which can be used to form an assessment about its quality, reliability, or trustworthiness. The PROV ontology, which is a W3C recommendation since 30th April 2013, provides definitions for interchange of provenance information. Using *PROV*, we can define entities and the various relations and operations between them such as generated by, derived from, and attributions. PROV has been successfully utilised in several

[1] https://www.w3.org/TR/owl2-overview/
[2] https://www.w3.org/TR/prov-o/
[3] purl.org/net/p-plan
[4] purl.org/adaptcentre/openscience/projects/CDMM

domains and applications[5] including encapsulation of scientific workflows [2,11] and provenance repositories [3,6] as well as in publication of experiment workflows. An *Entity* in *PROV-O* is defined as being physical, digital, conceptual, or other kind of thing with some fixed aspects. *PROV-O* defines an *Activity* as something that occurs over a period of time and acts upon or with entities; it may include consuming, processing, transforming, modifying, relocating, using, or generating entities.

PROV-O was designed to be generic and domain independent, and needs to be extended to address the requirements of representing consent and data lifecycles based on GDPR. Additionally, provenance in *PROV-O* only refers to executions that have already happened, but does not offer a vocabulary to express the 'plan' or 'template' that the execution was supposed to follow. *PROV-O* does contain the term *prov:Plan*, but the ontology itself does not elaborate on how the plans can be structured or used to relate to other provenance terms used in an execution.

P-Plan is an ontology that extends *PROV-O* to describe abstract scientific workflows as plans and link them to their past executions. A *p-plan:Plan* is a subclass of *prov:Plan* and is composed of smaller activities or steps (*p-plan:Step*) that use and produce (as inputs or outputs of steps) variables (*p-plan:Variable*). Together with the steps and variables, a *p-plan:Plan* represents provenance information of 'how' something should happen or a 'template' for executions. A *p-plan:Activity* is a subclass of *prov:Activity* and represents the execution of the process planned in a *p-plan:Step*. A *p-plan:Entity* is a subclass of *prov:Entity* that corresponds to a *p-plan:Variable* in the overall *p-plan:Plan*. Therefore, a *p-plan:Step* may describe the template including inputs and outputs which can then be instantiated into multiple instances of *p-plan:Activity* that can have distinct inputs to produce different outputs.

As *p-plan:Plan* extends *prov:Plan*, which itself extends *prov:Entity*, it can be used to treat the *p-plan:Plan* as an object whose provenance can be tracked using *PROV-O* or *P-Plan*. This makes it possible to express provenance of provenance, thereby creating a history of how activities and their interactions changed over time.

Extending provenance ontologies allows *GDPRov* to express a 'template' or 'plan' of what should happen (using *p-plan:Plan*) describing a model of all activities (as *p-plan:Step*) that can take place. This template is then instantiated for (using *p-plan:Activity*) each specific use of the activity, such as obtaining consent or data for a particular user. Additionally, the provenance of the activities themselves can be expressed (using *PROV-O* and *P-Plan*) to record how they change over time, making it possible to trace the change in activities along with how they interact with consent and data. This is beneficial in documenting the state of a system as a set of activities that deal with consent and data, and can be helpful in determining changes in consent when the interactions between data and an activity change over time. For example, differences in provenance of an activity can show that it uses personal data it did not previously use.

[5] https://www.w3.org/TR/prov-implementations/

Depending on the consent obtained for that particular user, this may or may not need additional permissions, and therefore require obtaining consent to permit such use. More information related to research regarding provenance traces in workflows can be found in [8, 9, 12].

3.2 Separation of consent and data processes

As a requirement under the GDPR, consent needs to be obtained before any collection, usage, or sharing of data can take place. To emphasise on this separation, *GDPRov* defines separate terms for 'data' and processes related with consent and personal data. This allows for simplification in modelling the two as each can reference the other without specifying their origin or history due to the open-world assumption in linked open data vocabularies. For example, processes that involves use of personal data can specify the consent that permits the usage without specifying how the consent was obtained or changed as the provenance metadata for the consent is defined separately. This distinction also allows activities to be divided or categorised in a modular manner, which can be helpful in representing internal organisational categorisation of these concepts.

3.3 Consent Agreement

Consent is defined by the GDPR [7] as "*any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she by statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her*". This can be represented as an agreement over three things - the personal data, usage of personal data, and the consent itself.

Consent is individually distinct for each user, but is common in terms of choices offered for all users. *GDPRov* represents this commonality of consent through subclasses of *p-plan:Variable* termed as *ConsentAgreement* and *ConsentAgreementTemplate*. *ConsentAgreement* is the consent obtained from the user, and reflects the choices made by the user which specify the permissions or restrictions for use of personal data. *ConsentAgreementTemplate* is the common template for the choices offered to all users regarding consent permissions and forms the basis on which the consent is acquired. An example would be a web form for obtaining consent, where *ConsentAgreementTemplate* would be the form consisting of fields and options as the common template of choices offered, and *ConsentAgreement* representing the choices made by the user as values submitted through the form. *GDPRov* also defines *TermsAndConditions* to reference the terms and conditions that are displayed when obtaining the consent or during activities such as registration.

A *ConsentStep* is a subclass of *p-plan:Step* that deals with consent and is further subclassed to distinguish between modification, acquisition, and archival of consent. Withdrawal of consent is considered a special case where the user revokes any previously given consent and is represented as a subclass of modification where the user revokes any previously granted permission over the same context.

Such an arrangement benefits semantics of change in consent where withdrawal is also visible in queries looking for modifications to the consent. As the *ConsentStep* is responsible for actions involving consent, entities based on consent such as *ConsentAgreement* and *ConsentAgreementTemplate* can only be acquired, changed, or deleted by a *ConsentStep*.

3.4 Data

GDPRov uses *Data* as a generic term to specify any data used in the provenance of steps and is further subclassed to represent personal data as *PersonalData*. *UID* is a subclass of Data that is used in profiling for referring to individual users. *AnonymousData* is a subclass of *PersonalData* and represents data that is pseudo-anonymous or anonymous. It's anonymity levels are represented using the object property *hasAnonymityLevel* which defines the degree to which it can be de-anonymised through defined instances of the class *AnonymityLevel*. Each instance of *AnonymityLevel* refers to the level of anonymisation (or the possibility of de-anonymisation) of the data instance, with *GDPRov* defining four instances to reflect the varying levels discussed in Section 2.5. The four instances, in increasing degree of anonymity are titled - *DeAnonymised*, *PseudoAnonymised*, *PseudoOrganisationalAnonymised*, and *Anonymised*. *GDPRov* only defines the anonymity level in abstract terms, and does not currently enforce any constraints on the use and declaration of anonymous data that can guide how processes act on it. This aspect may change in future versions of the ontology depending on the need for such mechanisms to reflect various use-cases.

A *DataStep* is a subclass of *P-Plan:Step* and represents steps that use or generate *Data*. *GDProv* specifies that only *DataStep*s may use *Data* objects so as to enable coherent queries that can retrieve all steps that use data in some capacity. *DataStep* is further subclassed to distinguish between collection (*DataCollectionStep*), deletion (*DataDeletionStep*), sharing (*DataSharingStep*), storage (*DataStorageStep*), and transformation (*DataTransformationStep*). *DataAnonymisationStep* is a subclass of *DataTransformationStep* and refers to the process where data is converted to a pseudo-anonymous or anonymous state reflected through the *hasAnonymityLevel* object property of the data object. Similarly, *DataArchivalStep* is a subclass of both *DataTransformationStep* and *DataStorageStep* as data undergoes transformation to some format before being stored in the form of an archive.

3.5 Process

GDPRov:Process is a subclass of *P-Plan:Plan* that combines a set of steps into a cohesive activity and can be used to reflect processes or services as a collection of steps that interact with data or consent. *GDPRov* defines certain subclasses of Process for GDPR mandated rights such as data erasure (*DataErasureProcess*), consent withdrawal (*ConsentWithdrawalProcess*), data rectification (*DataRectificationProcess*), data access (*DataAccessProcess*), and data archival (*DataArchivalProcess*).

These subclasses reflect the provenance trace for the series of actions that should be executed whenever an user exercises the particular right referred to by the process which outline the effects on data and consent used to comply with the rights being exercised. *HandleDataBreach* is a subclass of Process and is used to describe actions undertaken in the event of a data breach. GDPR requires notification to the Data Protection Office in 72 hours, along with notifying users about the impact of data breached [7]. Therefore, the steps under the *HandleDataBreach* process must reflect these activities through its series of steps.

3.6 Plans and Executions

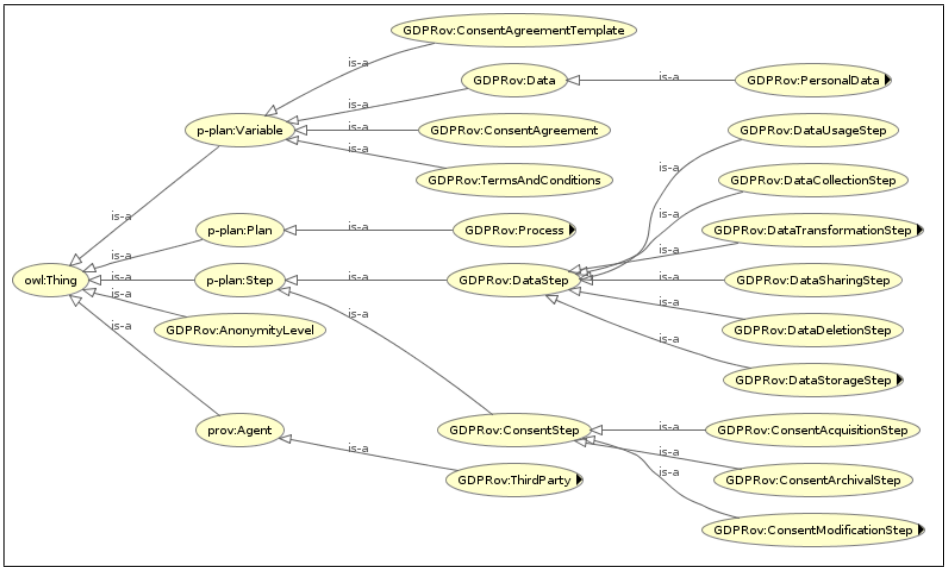

Fig. 1: GDPRov vocabulary hierarchy visualised using OWLViz/Protege

Using the *P-Plan* provenance model, *GDPRov* concepts describe a template of 'how' something should happen, which can be instantiated into concrete executions based on real-world usage. Fig.1 shows a partial hierarchy of the terms in *GDPRov*. Models based on the *P-Plan:Variable* such as *ConsentAgreement* and *Data* are instantiated as *P-Plan:Entity*, and are connected to the template through the *p-plan:correspondsToVariable* object property. Similarly, steps based on *P-Plan:Step* are instantiated as *P-Plan:Activity*, and use the *p-plan:correspondsToStep* object property.

Concepts such as *ConsentAgreementTemplate* which are common to all users need to be instantiated only once with links to this instance referenced in all uses. Concepts which are unique for each user, such as *ConsentAgreement*, will have a distinct instance for each user. In this case, even though the instance of the step

that obtains consent may be common for all users, the provenance metadata will be distinct for each user as it refers to the particular consent obtained from that user.

Because *P-Plan* provides links (via object properties) between the 'plan' and the 'execution', it is possible to query all instantiations of a particular step, or conversely, to query all activities that were undertaken based on a user's consent. This provides the opportunity to provide a log of activities carried out to create a documented proof of compliance and to demonstrate that the plan of activities follows the privacy by design model [14].

4 Compliance related queries using *SPARQL*

In this section, we discuss how *SPARQL* can complement *GDPRov* for formulating provenance based queries related to GDPR compliance. Provenance metadata, by itself, describes the lifecycles of the data and consent, which provide information about how they originated, how they were changed or modified, and usage by activities within their lifecycles. We discuss the use of *GDPRov* to retrieve this information through *SPARQL* queries using an example of a general shopping website.

The shopping website is an online-only service that allows registered users to purchase products sold in its marketplace. Information collected by the website includes user's billing and shipping details, which it claims is justified for shipment of the item, and therefore, is not under obligation to obtain an explicit consent. However, it can also store the user's shipping details and purchase history to save user's the effort of typing it for each order. The data from all purchases is shared with a third party for purposes of marketting, analytics, and targeted advertising. The data shared is in a pseudo-anonymised form so that they can de-anonymise the returned results, but at the same time, also not share personal data with the third party. The explicit consent obtained for both uses of data is via a web-form that lists the data activities, their justification, and a legal-text terms and condition outlining the services.

Using *GDPRov*, the above use case is modelled using *ConsentAgreementTemplate* to represent the web-form used to obtain the consent, which also includes *TermsAndConditions*, and where the user consent is obtained as *ConsentAgreement*. The data is shared with the third party using *DataSharingStep*, which shares an instance of Data called *AggregatedPurchaseData* which is anonymised using some *DataAnonymisationStep*.

A natural query over this data would be to retrieve information that can help determine whether the consent was correctly obtained, which requires retrieving all combinations of entities used in the collection of consent from users. This translates into the *SPARQL* query in Listing.1.1, where the tuples returned are of the form $(consent, template, T\&C)$ representing their combined usage in obtaining user consent.

Another example query retrieves all *Data* entities shared with third parties, the step responsible for the actual sharing, and whether they are anonymised along

```
PREFIX GDPRov:
<http://purl.org/adaptcentre/openscience/ontologies/gdprov#>}

SELECT ?consent ?template ?toc
WHERE {
        ?consent a GDPRov:ConsentAgreement .
        ?template a GDPRov:ConsentAgreementTemplate .
        ?toc a GDPRov:TermsAndConditions .
        ?step a GDPRov:ConsentAcquisitionStep .
        ?step GDPRov:usesConsentAgreementTemplate ?template .
        ?step GDPRov:usesTermsAndConditions ?toc .
        ?step GDPRov:generatesConsentAgreement ?consent
}
```

Listing 1.1: *SPARQL* query retrieving entities involved in acquiring user consent

with the anonymisation step responsible. The results of the query in Listing.1.2 are returned as a tuple $(data, sharestep, isAnonymised, anonymisationStep)$, shown in a simpler form in 1, where *isAnonymised* can be either *True* or *False*, and the value of *anonymisationStep* reflects the particular *DataAnonymisationStep* responsible for anonymisation.

The intentionally simple use-case demonstrates the use of *SPARQL* over *GDPRov* to retrieve information which provides helpful information about GDPR compliance. More complex queries and examples are possible involving OWL axioms and reasoning, for example, tracing origin of data, or generating history of consent associated with a particular user; but are out of scope for the current paper.

```
PREFIX GDPRov:
    <https://openscience.adaptcentre.ie/ontologies/GDPRov#>

SELECT ?data ?sharestep ?isAnonymised ?anonymisationStep
WHERE {
        ?data a GDPRov:Data .
        ?sharestep a GDPRov:DataSharingStep .
        ?sharestep GDPRov:sharesData ?data.
        BIND (
            EXISTS { ?data a GDPRov:AnonymisedData . }
            as ?isAnonymised ) .
        OPTIONAL {
                ?anonymisationStep
                GDPRov:generatesAnonymisedData ?data .
        }
}
```

Listing 1.2: *SPARQL* query retriving data shared with third parties

data	shareStep	isAnonymised	anonymiserStep
productsSold	productAnalytics	false	NULL
billingInfo	billingAnalytics	false	NULL
customerInfo	profiling	true	anonymiseUsers

Table 1: Results for SPARQL query in Listing. 1.2 for anonymity of shared data

5 Related Work

There have been several approaches related to expressing the GDPR as an ontology. [1] presents an ontology modelling data protection concepts using the terminology relevant to GDPR requirements targeted towards business processes. The ontology is described as a work-in-progress with a security related ontology described as potential for future work. Particularly of interest is the way it defines and links the concepts for *Consent* and *Compliance* using *Principles* such as *Fairness* and *Trust*. Though the ontology does not contain any provenance information, it can be used to augment *GDPRov* with legal terms for describing GDPR relevant concepts. However, in its current form, we found the ontology not suitable for use within GDPRov due to lack of provenance related concepts. In future, it may potentially be useful for describing compliance related information.

[16] demonstrates how *ODRL*[6] can be used to model and auto-generate access policies along with an enforcement framework for linked data markets. The authors discuss the distinction between enforceable and non-enforceable policies and use *ODRL* to auto-generate contracts for the latter based on a request mechanism. The use of such a mechanism in providing an agreement between data processors and controllers is of particular interest in lieu of GDPR. The auto-generation of contracts provides a better level of granularity in data sharing with third parties, which would result in better documentation of how data was shared between two parties. The proposed use of *ODRL* can be extended to complement *GDPRov* for expressing data sharing agreements with third parties.

The *UsablePrivacy*[7] project uses natural language processing and crowdsourcing for annotating website privacy policies. *PrivOnto* [13] is a semantic framework that uses SPARQL to query a corpus of 115 privacy policies and presents it in an interactive online tool. *PrivOnto* describes privacy policies in terms of fragments which can range from several words to sentences. It also contains terms for describing the annotation action itself by specifying the annotator and annotation relationship. A similar approach can be taken towards privacy policies that address the GDPR. However, as the GDPR is yet to come into effect, very few (if any) privacy policies address it. A converse approach could be creating a template addressing the GDPR and using it to map future privacy policies based on their implemented approach. *GDPRov* provides a suitable model to express the use of personal data mentioned in privacy policies that can be queried to

[6] https://www.w3.org/ns/odrl/2/
[7] http://usableprivacy.org/

form similar dashboards with actual use of the data, possibly through automated mechanisms.

PrivacyInsight [4] is a privacy dashboard that maps data to information flows using provenance information and targets rights granted under the GDPR. Information flow is presented as a formal model of acyclic provenance graphs where the root of the graph depicts the source of the data. The graph is then visualised to provide a representation of information flows in the system which the user can interact with. Provenance is collected from multiple systems and only aggregated upon request by the data subject. Event listeners and embedded in and collect provenance information from processing layers such as operating system, applications, and databases. Of particular interest is the visualisation of provenance traces based on collected information, which can also be applied to visualise GDPRov use-cases.

GDPR not only affects commercial organisations, but also applies to user data collected for research by universities and research institutes as discussed by [10] and [5]. The intended use of data in experiments needs particular focus as data collected as part of one experiment may not be eligible under the obtained consent for use in future experiments. Generally, in academia, user consent is collected in the form of a verbose agreement in print or digital form and is accompanied by a description of the experiment. The consent form contains information about the data collected, its intended use (in the form of experiment description), and the possibility of publication, in which case it also states whether the data will be anonymised. An alternative to this approach is to maintain records of the consent containing the specific permissions regarding collection and usage of data along with the provenance of intended activities using an ontology such as *GDPRov*. This allows the linking of obtained consent with its usage in publication and can help with determining conditions for potential future use. It also allows the published dataset to hold the consent obtained from the user along with a provenance description how the consent and data were obtained and processed.

6 Conclusion & Future Work

With the GDPR set to come into effect on May 25$^{\text{th}}$ 2018, it is important to look into tools and technologies which can help with its compliance. Transparency is a key factor in demonstrating compliance as it allows all parties involved to determine whether the consent and data are/were obtained and used fairly and correctly. In this paper, we present *GDPRov*, a linked data ontology to describe the provenance of activities such as acquisition, usage, storage, deletion, and sharing of consent and data lifecycles. *GDPRov* extends *PROV-O*, which is a W3C recommendation, and *P-Plan* to describe provenance metadata for what is supposed to happen and executions showing what has taken place. The ontology provides terms to describe various levels of anonymisation as GDPR stipulates different obligations depending on how the data can be de-anonymised. *GDPRov* also provides terms to expresses the obligations of handling various rights such as consent modification and withdrawal, and requesting rectification or

access to data. The paper discusses the use of *SPARQL* to query the provenance information described by *GDPRov* and its use in identifying information relevant for compliance. *GDPRov* also provides a way to describe what steps will be taken for certain activities mandated by the GDPR such as reporting a data breach.

GDPRov and the work described in this paper are part of a larger ongoing project for a consent management framework. *GDPRov* is a therefore work-in-progress, and as such is not yet final or complete in terms of being suitable to reflect all the actions stipulated under the GDPR. In terms of future work, we will be working on refining the ontology through adoption of various use-cases based on compliance expectations. To describe the consent using a suitable ontology, we are evaluating the use of *ODRL* and *XACML* to express the agreement between the user and the service provider, and to formalise it as a set of agreements that can be queried.

Data described by *GDPRov* provides a large area of opportunity for future work in terms of describing compliance as a set of prospective *SPARQL* queries. One such use case could be provenance metadata describing how activities change over time in their use of data which can be used to determine if new consent needs to be obtained from the user. Additionally, the queries themselves can be recorded or captured using ontologies such as *SWRL*[8] and *SPIN*[9], which makes it possible to record the use of such queries as documentation for a compliance tool. Augmenting the GDPRov metadata with *SHACL*[10] can help express constraints over the use of data by activities, which can be extended to express them as compliance-related obligations. Such approaches using linked open data can prove helpful in determining whether changes proposed in the provenance will be compliant and to highlight areas that need attention.

It would be helpful to have the GDPR text as a referenceable resource using linked open data for referencing the appropriate legal text as helpful documentation in relation to compliance. We are in the process of creating such a resource (available online[11]), and in future expect to annotate *GDPRov* and the related *SPARQL* queries with references to the legal text based on this work.

Acknowledgement

This paper is supported by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

[8] https://www.w3.org/Submission/SWRL/
[9] https://www.w3.org/Submission/spin-overview/
[10] https://www.w3.org/TR/shacl/
[11] purl.org/adaptcentre/openscience/projects/GDPRtEXT

References

1. Bartolini, C., Muthuri, R., Cristiana, S.: Using ontologies to model data protection requirements in workflows. In: Ninth International Workshop on Juris-informatics (JURISIN 2015). http://orbilu.uni.lu/handle/10993/22383 (2015)
2. Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., Palma, R., Mina, E., Corcho, O., Gómez-Pérez, J.M., Bechhofer, S., et al.: Using a suite of ontologies for preserving workflow-centric research objects. Web Semantics: Science, Services and Agents on the World Wide Web 32, 16–42 (2015)
3. Belhajjame, K., Zhao, J., Garijo, D., Garrido, A., Soiland-Reyes, S., Alper, P., Corcho, O.: A workflow prov-corpus based on taverna and wings. In: Proceedings of the Joint EDBT/ICDT 2013 Workshops. pp. 331–332. ACM (2013)
4. Bier, C., Kühne, K., Beyerer, J.: Privacyinsight: the next generation privacy dashboard. In: Annual Privacy Forum. pp. 135–152. Springer (2016)
5. Chassang, G.: The impact of the eu general data protection regulation on scientific research. ecancermedicalscience 11 (2017)
6. Cuevas-Vicenttín, V., Kianmajd, P., Ludäscher, B., Missier, P., Chirigati, F., Wei, Y., Koop, D., Dey, S.: The pbase scientific workflow provenance repository. International Journal of Digital Curation 9(2), 28–38 (2014)
7. 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). Official Journal of the European Union L119, 1–88 (May 2016), http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC
8. Garijo, D., Corcho, O., Gil, Y., Gutman, B.A., Dinov, I.D., Thompson, P., Toga, A.W.: Fragflow automated fragment detection in scientific workflows. In: e-Science (e-Science), 2014 IEEE 10th International Conference on. vol. 1, pp. 281–289. IEEE (2014)
9. Koohi-Var, T., Zahedi, M.: Linear merging reduction: A workflow diagram simplification method. In: Information and Knowledge Technology (IKT), 2016 Eighth International Conference on. pp. 105–110. IEEE (2016)
10. Lewis, D., Moorkens, J., Fatema, K.: Integrating the management of personal data protection and open science with research ethics. In: Ethics in NLP Workshop, EACL. pp. 60–65 (2017)
11. Missier, P., Dey, S.C., Belhajjame, K., Cuevas-Vicenttín, V., Ludäscher, B.: D-prov: Extending the prov provenance model with workflow structure. In: TaPP (2013)
12. Missier, P., Woodman, S., Hiden, H., Watson, P.: Provenance and data differencing for workflow reproducibility analysis. Concurrency and Computation: Practice and Experience 28(4), 995–1015 (2016)
13. Oltramari, A., Piraviperumal, D., Schaub, F., Wilson, S., Cherivirala, S., Norton, T.B., Russell, N.C., Story, P., Reidenberg, J., Sadeh, N.: Privonto: A semantic framework for the analysis of privacy policies. Semantic Web (Preprint), 1–19 (2016)
14. Schaar, P.: Privacy by design. Identity in the Information Society 3(2), 267–274 (2010)
15. Schwartz, P.M., Solove, D.J.: Pii 2.0: Privacy and a new approach to personal information. Bloomberg BNA: Privacy & Security Law Report, 11 PVLR 142 (2012)
16. Steyskal, S., Kirrane, S.: If you can't enforce it, contract it: Enforceability in policy-driven (linked) data markets. In: SEMANTiCS (Posters & Demos). pp. 63–66 (2015)