PRIVACY AND LINKED DATA ON THE INTERNET – WHAT TO ANNOTATE AND HOW?

Miroslav Vacura

Faculty of Economics
Prague University of Economics and Business
vacuram@vse.cz

Keywords

Privacy, linked data, vocabulary, dataset, annotation

Abstract

The emergence of a growing number of lightweight vocabularies, in addition to the large number of datasets based on them, has increased the need to discuss potential privacy challenges in the context of Linked Data on the Internet. While existing research has identified a number of potential privacy issues and proposed some tailored solutions for them, we still lack a consensus for dealing with privacy challenges within Linked Data in general. In this paper we present two commonly used ways of handling privacy-sensitive data in a Linked Data environment: ontologies for representing privacy, and vocabularies for annotating of individual datasets. Then we proceed to propose a novel alternative approach: annotating individual vocabularies in their entirety, and annotating individual elements of those vocabularies.

1. Introduction

The Linked Data initiative was started by Berners-Lee (2006) as an innovative approach to the concept of the Semantic Web. He understood the idea of the Semantic Web as a web of data explicitly linked across different datasets, enabling both people and machines to integrate and explore them semantically – that is, with specific regard to meaning. Linking data implies that access to a piece of data enables a user to find pieces of related data iteratively. In the same way that standard HTML web hyperlinks relate documents to each other, linked data connects *entities* by using RDF formal language. Berners-Lee (2006) defined the following requirements for Linked Data:

- Use URIs as names for things.
- Use HTTP URIs so people can look up those names.
- When someone looks up a URI, provide useful information, using standards (RDF, SPARQL, OWL etc).
- Include links to other URIs, so that people can discover more things.

Those general principles provide only guidance for future development. They obviously lack any mention of privacy; in fact the issue of privacy was altogether neglected during the enthusiastic formative years of fulfilling this vision. Recent papers do however deal with specific privacy issues and attacks in the Linked Data environment (Heitmann et al., 2016; Miracle and Cheatham, 2016).

This paper discusses some more general approaches to protecting privacy within Linked Data technologies.

The rest of the paper is organized as follows: Section 2 provides basic characterization of the Linked Data project, its history, and associated standards and technologies. Section 3 provides the theoretical and legal motivation for privacy issues discussed in this paper. Section 4 presents two common ways of handling privacy issues in the Linked Data environment: ontologies for representing privacy, and vocabularies for annotating individual datasets. Then we proceed to propose two innovative alternative approaches: annotating vocabularies as a whole and annotation of individual elements of those vocabularies. Finally, Section 5 provides some conclusions.

2. Linked Data

Initial guiding principles of Linked Data as described in the previous section were later fully developed in a number of technical documents (Bizer et al., 2007; Sauermann et al., 2007) and overview papers (Bizer et al., 2008; Bizer et al., 2009).

Linked data can be queried e.g. through the formal query language SPARQL, which can be used to retrieve information in similar fashion to relational databases. Linked data can also be crawled with appropriate browsers, in a way comparable to usual web pages, by following RDF links. However, while HTML provides only a generic linking capability, the Linked Data environment enables the designer to create semantically different types of links: we can, for example, specify that an individual person is *author of* an individual paper – and instead of making a general link between person and paper it determines the specific type of link: *author of*. There are also search engines that are able to retrieve RDF information triplets and search across the whole universe of Linked Data, that at time of writing (January 2021) comprises more than 1,255 datasets with 16,174 links (see LOD cloud[18]). These datasets include some of the most popular data sources available on the internet, such as DBpedia that describes about 6 million entities.

The Linked Open Vocabularies[19] (Vandenbussche et al., 2017) initiative is an 'observatory' of the ecosystem of semantic vocabularies, started in March 2011 as part of the DataLift research project (Scharffe et al., 2012) and hosted by the Open Knowledge Foundation. To date it has registered more than 600 different mostly lightweight vocabularies.

Linked Data was built without any explicit regard for privacy and therefore it is in critical need of a review specifically targeted to identify potential privacy issues. Ontology schemata such as the DBpedia ontology and schema.org have been developed without marked regard for privacy and its enforcement in the Linked Data environment. Some vocabularies like *Friend of a Friend* (FOAF) explicitly focus on personal information and interpersonal relationships, and yet they do not have any features to handle privacy-sensitive information; thus, they may represent an extensive privacy risk (Vacura et al., 2016).

After the success of semantic technologies in the industry in recent years based on widespread use of lightweight vocabularies, it can be argued that a system of good practices, ensuring that all legitimate privacy concerns are being reflected and handled properly, is long overdue.

The key problem illustrated above is the vast amount of existing datasets and vocabularies. Discussion surrounding privacy issues has thus far been overlooked, so we already have to consider hundreds of vocabularies and thousands of datasets with no relevant information about possible

[18] https://lod-cloud.net/

[19] http://lov.okfn.org/dataset/lov/

privacy issues connected to their use. In following discussion we analyze whether currently available solutions are viable and propose some new approaches.

3. Privacy in the Context of Complex Regulatory Framework

Any approach to annotation of linked data with regards to privacy issues has to take into account relevant EU regulatory frameworks. In this section we provide brief theoretical and legal context for privacy issues discussed in this paper, although we will not dive into details. The purpose of this section is to emphasize complexity of these considerations and pinpoint key terms, that are relevant for our projects.

The starting point for our discussion about privacy is Article 8 of the *EU Charter of Fundamental Rights*[20] that in general terms declares protection of personal data as one of our fundamental rights. Specifically, it is concerned with fairness and purposefulness of personal data processing, the right to access personal data, the requirement of consent of the person concerned or other legitimate basis written in law to process personal data and so forth. Those principles were recently further extended by the General Data Protection Regulation (GDPR).[21]

Alongside the concepts and regulations defined by GDPR are some notions closely related to them. *Relational Law* is understood as the intertwined complex of regulatory frameworks of law and programming (Casanovas, 2013). Regulatory *system* relates to the social concerns of relational law as well as procedural ways to manage and solve conflicts. *Relational justice* is shaped by regulatory systems and is a type of justice emerging from specific practices and strategies within technological situated contexts (Casanovas and Poblet, 2008).

Regulatory Models (RM) comprise of a normative suite implemented in computing environments that monitors compliance with regulatory systems and relational justice, the specific structure of principles, values, norms, and rules guiding technical protocols, multi-layered relation of organizations (multi-layered governance), and the interoperability of computer languages. In a case where such a regulatory model is based on semantic technologies, the term *Semantic Web Regulatory Model* (SWRM) has been introduced (Casanovas, 2015).

SWRMs can be integrated within a larger framework under the meta-rule of law in order to facilitate their use, e.g. the requirements of the *European Market Infrastructure Regulation*[22] as demonstrated by Casanovas et al. (2016).

The list of privacy challenges that is to be regulated and handled by SWRMs is extensive, and relative to applicable fields. Analysis and the list of privacy concerns can be found in Iacob and Bikakis (2016). Different approaches to annotating semantic structures as proposed in the following sections of this paper are envisioned to serve as a technical basis for future specifications of SWRMs, thus potentially assisting further implementation of the above-mentioned legal norms within real-world technology.

[20] http://www.europarl.europa.eu/charter/pdf/text_en.pdf

[21] https://ec.europa.eu/info/law/law-topic/data-protection_en

[22] http://ec.europa.eu/finance/financial-markets/index_en.htm

4. Privacy Concerns and Light-weight Vocabularies

Complex theoretical and technological frameworks mentioned in the previous section rely on the ability to formally describe entities relevant to privacy concerns and facilitate automatic or semi-automatic management of privacy and data protection during instances of data transfer and data exchange.

Semantic Web technologies provide several approaches to semantically describe these entities and their relevance to privacy-related concerns. These approaches however differ in a number of respects and these differences influence their practical usability and effectiveness. However, when using Semantic Web (SW) tools, *Regulatory Model* (RM) turns into *Semantic Web Regulatory Model* (SWRM) as described in the previous section.

In following sections we firstly discuss two commonly used approaches: ontologies for representing privacy and annotations of individual datasets. Then we proceed to introduce two novel alternative approaches: annotating whole vocabularies and annotations of individual elements of vocabularies.

4.1. **Ontologies for representing privacy**

Privacy ontologies can be divided into *general* and *specialized* based on their intended use. General privacy ontologies are domain-independent and can be used in any scenario or context. They are not connected to any context-specific concepts or notions. Specialized ontologies focus on a specific domain or area and are reliably usable only within the boundaries of their intended coverage.

The most widely used ontologies for representing privacy today are specialized and focus only on a specific domain, such as IoT (Das et al., 2016). The terminology and structure of these ontologies reflects the area of their intended use. The main advantage of using them is that such ontologies are custom-made for a particular use, therefore, they are typically more effective and practical. A notable drawback is that their use is possible only within those boundaries, so their use in the context of Linked Data is limited to vocabularies related to domains covered by these ontologies.

An example of specialized privacy ontology is *HL7 Security and Privacy Ontology*, which serves to name, define, formally describe, and interrelate key security and privacy concepts within the scope of Healthcare Information Technology.[23]

A brief overview and evaluation of other specialized ontologies for representing privacy can be found in Casanovas et al. (2016) and Iacob and Bikakis (2016). The authors of the first of these papers however conclude that while those ontologies are very often comprehensively designed in terms of considering theoretical aspects, they are in practice difficult to learn and use.

An example of a general privacy ontology relevant to our discussion is *Privacy Preference Ontology* (PPO) for Linked Data (Sacco, 2011), that builds on the *Web Access Control* (WAC) vocabulary.[24] PPO restricts RDF documents to provide fine-grained privacy measures to specify access restrictions to the data. This example illustrates a common limitation of general ontologies – although such ontologies are not limited to a single domain, they usually focus on a single aspect of privacy, e.g. access rights. Based on our ongoing research we have found no single comprehensive general privacy ontology that would cover all aspects of privacy.

[23] http://wiki.hl7.org/index.php?title=Security_and_Privacy_Ontology

[24] https://www.w3.org/wiki/WebAccessControl

These ontologies also focus on annotating individual data pieces (Abox): representing a privacy preference to share an individual e-mail address of one person that may be different than the preference of another person (PPO).

There is also another significant drawback of using existing comprehensive privacy ontologies in a Linked Data environment. Ontologies with heavyweight axiomatization may be useful for tasks that require complex reasoning. However, in a Linked Data context typically lightweight vocabularies are used and consequently only very limited axiomatization and reasoning is employed, if it is employed at all. Complex ontologies do not fit seamlessly into the Linked Data environment and sometimes even present a significant obstacle. Our brief overview shows that there is currently no general privacy ontology that could provide a technical basis for future specifications of SWRMs. If it existed, it could also serve as a starting point for approaches described in following sections.

Another problem stems from the amount of existing unannotated datasets mentioned in Section 2. Manual or semi-automatic interlinking of elements of those datasets with a privacy ontology is inconceivable. It may be possible to implement some form of automatic ontology mapping, but it is not clear whether this is feasible given the sensibility of privacy information and implied requirements for precision.

4.2. **Annotating individual datasets**

Another approach to privacy in Linked Data comprises annotation of datasets. As an example of this approach we propose the use of well-known standard *Dublin Core* together with *Linked Data Rights* (LDR)[25] vocabulary (Rodríguez-Doncel, 2013). The LDR vocabulary enables the user to create policies and expressions concerning rights for Linked Data resources. LDR focuses on intellectual property rights and makes it possible to say that some resource (ldr:Resource) has some legal status (ldr:ResourceLegalStatus) etc.

Dublin Core combined with LDR can be used to describe privacy information related to the dataset, however, the scope of privacy issues it covers is very limited. There are further drawbacks mentioned by Casanovas et al. (2016), for example some missing properties essential to effective utilization of this vocabulary. Those properties include the specific country where the personal data file has been registered, the privacy level of the dataset, and security measures that should be taken. Overall, available tools for annotating privacy of individual datasets are limited; there is a clear need for them to be extended and further interconnected.

A similar problem as noted in the previous section is related to the amount of existing unannotated datasets. Manual or semi-automatic annotation of those datasets is again inconceivable. It may be possible to implement some form of automatic annotation, but again this may not be feasible given the sensibility of privacy information.

We can also observe that there is currently no way to properly annotate individual datasets with regards to privacy that could provide a technical basis for future specifications of SWRMs

4.3. **Annotating individual vocabularies**

An alternative to the annotation of datasets is the annotation of whole vocabularies. The broad notion here is that while some vocabularies are typically used to handle privacy-sensitive information (e.g. *Friend of a Friend* – FOAF), other vocabularies may be marked as generally safe from privacy concerns (e.g. *Biological Taxonomy Vocabulary* – Botany). At the most general level, annotation would distinguish between safe and unsafe vocabularies, at a more detailed level it could

[25] http://purl.oclc.org/NET/ldr/ns

distinguish between various types of privacy concerns that different vocabularies raise. So, while existing ontologies such as PPO assign annotations to individual pieces of data (concrete e-mail address of a person), in this case we will assign annotations to a vocabulary as whole. E.g. the FOAF vocabulary (identified by its URI) would be annotated as "unsafe" – optionally with some other fine-grained privacy related attributes; however, still related to the vocabulary as whole, while the Botany vocabulary would be annotated as "safe" regarding privacy.

Our research to date suggests that there is no ontology or vocabulary that is well suited for this method of annotation of vocabularies. However, considering the straightforward nature of such a vocabulary we believe that its development would not be an unduly difficult or time-consuming task. There could be some formal or syntactical technical difficulties that would need to be resolved, related to the fact that the entity to be annotated is the whole vocabulary, in which case the annotation would need to be attached to the respective namespace identifying that vocabulary. Given the number of vocabularies discussed in Section 2 it may be an extensive task but we believe that it is still plausible to consider the future of semi-automatic or even manual annotation. Annotations of vocabularies could also provide a technical basis for future specifications of relevant privacy SWRMs.

Annotation of whole vocabularies may also be regarded as a first step to more detailed annotation of individual elements of those vocabularies, as discussed in the following section. Exhaustive annotation of vocabulary elements would only be necessary for those vocabularies that as a whole handle some privacy-relevant data.

4.4. **Annotating individual elements of vocabularies**

After we have identified and annotated a vocabulary that may be used to describe privacy-sensitive information as described in the previous section it would be useful to analyze elements of this vocabulary separately and annotate them individually. While the total number of vocabularies is high, it should be possible to use annotations described in the previous section to select and annotate only those vocabularies that are used to handle privacy-sensitive information.

As in the case of annotating whole dictionaries, there might be some syntactical difficulties, correlating with our use of one vocabulary to annotate individual elements of other vocabulary. Nevertheless, given the nature of the RDF standard such annotation should be feasible.

There is currently no vocabulary comprehensive enough to make such annotation possible. However, there is a considerable number of specialized vocabularies that can be extended or complemented by new vocabularies to facilitate such an annotative project. Nonetheless, such a project would by necessity be more complex and time-consuming than the approach mentioned in previous section.

5. Conclusions

This paper explores the possible ways of dealing with privacy challenges in the Linked Data environment. The emergence of lightweight vocabularies and the large number of datasets based on them has increased the need to discuss potential privacy challenges. Existing research has already identified a number of potential privacy issues and proposed some custom-tailored solutions for them (Heitmann, 2016; Miracle, 2016).

In this paper we first provided a characterization of the Linked Data initiative then discussed some concepts of the theoretical and legal backgrounds of privacy issues relevant to our topic. The following section of the paper consisted of a presentation of two commonly used ways of handling

privacy-sensitive data in a Linked Data environment: ontologies for representing privacy, and vocabulary annotation of individual datasets. We proceeded to propose a novel alternative approach: annotation of whole individual vocabularies and annotating individual elements of those vocabularies. These annotations could also provide a technical basis for future specifications of relevant privacy SWRMs, the importance of which was described in Section 3.

In future work on this topic, we plan to fully develop the annotation framework foreshadowed in Section 4.4. Our aim is to develop a vocabulary that can be used for annotating other vocabularies, specifically regarding their privacy status.

6. Acknowledgement

This work has been supported by CSF 18-23964S.

7. References

Berners-Lee, T. (2006). Linked Data. https://www.w3.org/DesignIssues/LinkedData.html. Document was updated on 2009. Accessed on 10th March 2021.

Bizer, Ch., Cyganiak, R. & Heath, T. (2007). How to publish linked data on the web. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/ Accessed on 10th March 2021.

Bizer, Ch., Heath, T., Idehen, K. & Berners-Lee, T. (2008). Linked data on the web. In Huai et al. Proceedings of the 17th International Conference on World Wide Web (WWW 2008), Beijing, China, April 21-25, 2008, 1265-1266.

Bizer, Ch., Heath, T. & Berners-Lee, T. (2009). Linked data the story so far. In International Journal on Semantic Web and Information Systems. 5 (3), 2009.

Casanovas, P., Poblet, M. (2008). Concepts and fields of relational justice. In P. Casanovas, G. Sartor, N. Casellas, R. Rubino (eds.), Computable Models of the Law. Languages, Dialogues, Games, Ontologies, LNAI 4884, Heidelberg, Berlin: Springer Verlag, 323-335.

Casanovas, P. (2013). Agreement and relational justice: a perspective from philosophy and sociology of law. In Sascha Ossowski (ed.), Agreement Technologies, LGTS n. 8, Dordrecht, Heidelberg: Springer Verlag, 19–42.

Casanovas, P., (2015). Semantic Web Regulatory Models. Why Ethics matter? Philosophy & Technology. 28(1), 33-55.

Casanovas, P., et al (2016). A European Framework for Regulating Data and Metadata Markets. In Brewster, Ch. et al. (ed.) Proceedings of the 4th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2016), co-located with 15th International Semantic Web Conference (ISWC 2016)}, Kobe, Japan, October 18th, 2016.

Das, P. K., et al (2016). Semantic Knowledge and Privacy in the Physical Web. In Brewster, Ch. et al. (ed.) Proceedings of the 4th Workshop on Society, Privacy and the Semantic Web – Policy and Technology (PrivOn2016), co-located with ISWC 2016, Kobe, Japan, October 18th, 2016.

Miracle, J., Cheatham, M. (2016). Semantic Web Enabled Record Linkage Attacks on Anonymized Data. In Brewster, Ch. et al. (ed.). Proceedings of the 4th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2016), co-located with ISWC 2016, Kobe, Japan, October 18th, 2016.

Heitmann, B., Hermsen, F., Decker, S. (2016). Towards the Use of Graph Summaries for Privacy Enhancing Release and Querying of Linked Data. In Brewster, Ch. et al. (ed.). Proceedings of the 4th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2016), co-located ISWC 2016, Kobe, Japan, October 18th, 2016.

Iacob, S., Bikakis, A. (2016) Evaluation of Semantic Web Ontologies for Privacy Modelling in Smart Home Environments. In Brewster, Ch. et al. (ed.) Proceedings of the 4th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2016), co-located with 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 18th, 2016.

Rodríguez-Doncel, V., Gómez-Pérez, A., Mihindukulasooriya, N. (2013). Rights declaration in Linked Data, in Proceedings of the 3rd Int. W. on Consuming Linked Data, O. Hartig et al. (eds) CEUR Workshop Proceedings Vol-1034.

Sacco, O., Passant, A. (2011). A Privacy Preference Ontology (PPO) for Linked Data. In Proceedings of Linked Data on the Web (LDOW2011), Hyderabad, India, March 29, 2011, CEUR Workshop Proceedings, Vol-813.

Sauermann, L., Cyganiak, R., Vlkel, M. (2007). Cool URIs for the semantic web. Technical Memo TM-07-01, DFKI GmbH, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany.

Scharffe, F., et al. (2012) Enabling linked-data publication with the datalift platform. In Hoffmann, J., Selman, B. (eds.). 26th Conference on Artificial Intelligence (AAAI-12). AAAI Press.

Vacura, M., Svátek, V., Gangemi, A. (2016). An ontological investigation over human relations in linked data. Applied Ontology, 11(3), 227-254.

Vandenbussche, P.-Y., et al (2017). Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. Semantic Web, 10(1), 437-452.