Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

Frequent Use Cases Extraction from Legal Texts in the Data Protection Domain

Valentina LEONE [1] and Luigi DI CARO
Computer Science Department, University of Turin, Italy

Abstract. Because of the recent entry into force of the General Data Protection Regulation (GDPR), a growing of documents issued by the European Union institutions and authorities often mention and discuss various use cases to be handled to comply with GDPR principles. This contribution addresses the problem of extracting recurrent use cases from legal documents belonging to the data protection domain by exploiting existing Ontology Design Patterns (ODPs). An analysis of ODPs that could be looked for inside data protection related documents is provided. Moreover, a first insight on how Natural Language Processing techniques could be exploited to identify recurrent ODPs from legal texts is presented. Thus, the proposed approach aims to identify standard use cases in the data protection field at EU level to promote the reuse of existing formalisations of knowledge.

Keywords. legal ontologies, ontology design patterns, NLP for legal texts

1. Introduction

Written documents are produced in every legal domain in order to spread the law. In the data protection domain, because of the entry into force of the General Data Protection Regulation (GDPR) on May 25th 2018, the debate about how to guarantee the protection of personal data has acquired a pivotal focus. The GDPR sets several measures and practises that different stakeholders dealing with the processing of personal data should adopt to protect data subject's rights and achieve a full compliance with the Regulation. These obligations and rules represent a set of use cases to be properly handled.

The need for the involved actors to comply with the new principles prescribed by the GDPR encouraged the modelling of computational models to support the automatic compliance checking. GDPRov [1], GDPRtEXT [2] and PrOnto [3] ontologies are the main examples of this effort. However, despite these resources model similar use cases, each of them adopts its own ontological commitment, i.e. its own perspective about the data protection domain. These different perspectives bring to ontological representations that, despite being characterised by some distinctive representational choices, share some similarities in the way in which they model the knowledge related to the field of interest.

The problem of redundant representations of knowledge clashes with the principles of reuse and economy of information promoted by the Linked Data [4] in the Semantic

[1] Corresponding Author: Valentina Leone. E-mail: leone@di.unito.it.

The work was partially supported by EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 690974 (MIREL).

Web context. Following this trend, Ontology Design Patterns (ODPs) were proposed as modelling solutions to solve recurrent ontology design problems [5].

In light of those considerations, this contribution addresses the problem of identifying, inside legal texts related to the data protection domain, the use cases for which a standardised modelling solution is already provided by an existing ODP. The approach relies on Natural Language Processing (NLP) techniques to automatically extract evidences of those patterns inside a corpus legal documents.

The paper is organised as follows: Section 2 presents some related works, Section 3 provides an overview of the ODPs that were selected to represent the data protection domain, Section 4 describes a preliminary experiment aimed at extracting one of the selected ODPs from legal documents through NLP, Section 5 ends the paper with the conclusion and the future work.

2. Related work

Legal ontologies in the data protection field. The Data Protection ontology[2] [6] was the first effort to provide a representation of the data protection domain including GDPR related concepts. More recently, GDPRov[3] [1] described the provenance of consent and the data life-cycle modelling abstract workflows to depict how consent and data are collected, used, stored, deleted and shared. GDPRtEXT[4] (GDPR text EXTension) [2] represents the relevant concepts expressed by the GDPR linking them to the parts of the Regulation containing the corresponding definitions. Finally, PrOnto (Privacy Ontology) [3,7] groups the concepts it represents in six macro-classes (i.e., personal data, rights and obligations, processing operations, roles, legal bases, purposes) and aims to provide a model on which approaches of legal reasoning and compliance checking can be applied.

Ontology Design Patterns. Ontology Design Patterns (ODPs) are small ontologies modelled as reusable components that provide a standardised representation of recurrent ontology design problems [5]. This definition implies the presence of use cases which occur frequently inside the domain of interest to be formally represented. A use case is usually expressed by formulating some competency questions for which the proposed ODP should be able to provide a modelling solution, making clear which are the involved entities and the interactions among them. Over the years, the Ontology Design Patterns Portal[5] [8] collected several contributions aimed to provide standardised solutions to different use cases, thus becoming the main reference on the Web for disclosing new ODPs.

Open Information Extraction. Open Information Extraction (OIE) [9] focuses on the extraction of <subject, predicate, object> triples from unstructured texts. Reverb [10] and DefIE [11] are some of the main contributions to OIE, the former adopting syntactical constraints, the latter applying a Word Sense Disambiguation step in order to filter out uninformative relations. Other approaches to OIE, such as KrankeN [12] and ClausIE [13] focus on the extraction of N-ary relations to address the loss of information resulting from limiting the extraction of triples to those identifying binary relations.

[2] http://bit.ly/2uhumDv

[3] https://openscience.adaptcentre.ie/ontologies/GDPRov/docs/ontology

[4] http://bit.ly/2xwjTZJ

[5] http://ontologydesignpatterns.org

Table 1. The list of CPs that were selected from the Ontology Design Pattern Portal and that model use cases of interest in the data protection domain.

Acting For	Action	Activity Specification
Agent Role	Complaint Design Pattern	Communication Event
Information Realization	Object Role	Part Of
Participation	Periodic Interval	Privacy Policy Personal Data
Task Execution	Time Indexed Participation	Time Indexed Person Role
Time Indexed Part Of	Time Indexed Situation	Time Interval
Time Period		

3. ODPs for the legal domain

A preliminary analysis of the Ontology Design Patterns Portal was performed in order to select candidate ODPs modelling use cases that could be possibly find in the data protection domain. In particular, the analysis focused on content design patterns (CPs) listed in the dedicated Web page[6]. CPs differ form other ODPs because the solutions they propose focus on the modelling of classes and properties of a domain, instead of providing domain-independent solutions more focused on solving design expressivity problems [14,15] .

The portal does not set constraints to the type of CPs that can be submitted, allowing to insert both patterns referring to a specific domain as well as patterns modelling general cross-domains use cases. A list of domains that can be associated to the CPs is provided by the portal and each pattern usually states the name of one or more domains it refers to. The selection of the CPs of interest, out of the 157 patterns listed in the portal, was performed analysing the competency questions associated to each pattern and evaluating its suitability for the data protection domain. As this domain is a multidisciplinary field that involves also the management of workflows, the scheduling of tasks and the handling of some events, the selected CPs do not only belong to the law field, but also to other different related domains (e.g. Management, Scheduling, Organization and Event Processing). Moreover, several patterns belonging to the General domain (i.e. patterns not specialised or limited to a range of subjects) were included. Table 1 shows the list of patterns that were selected after this analysis.

Among the selected patterns, only two of them are strictly related to the legal domain, i.e. the Complaint Design Pattern[7] [16] and the Privacy Policy Personal Data pattern[8] [17] . While the former allows the modelling of the different constituents found commonly in a complaint, the latter allows the representation of the information contained into a privacy policy describing how the personal data are processed.

Different groups of CPs can be identified considering the similarities holding among the use cases they model. For instance, some of the CPs focus on the modelling of a situation in which an agent (intended as a human being) is involved. By contrast, other CPs try to represent actions and events that require the modelling of temporal parameters. Table 2 shows a possible organisation of the CPs of interest according to different criteria.

[6] http://ontologydesignpatterns.org/wiki/Submissions:ContentOPs

[7] http://ontologydesignpatterns.org/wiki/Submissions:Complaint_Design_Pattern

[8] http://ontologydesignpatterns.org/wiki/Submissions:PrivacyPolicyPersonalData

Table 2. A list of CPs representing agents involved in some situation (left), a list of CPs representing actions and events involving the modelling of temporal aspects (centre) and a list of CPs related to the law field (right). Some of the CPs could appear in more than one column.

Agents	**Actions and events**	**Law field**
Acting For	Activity Specification	Complaint Design Pattern
Agent Role	Action	Privacy Policy Personal Data
Complaint Design Pattern	Communication Event	
Part Of	Participation	
Participation	Time Indexed Participation	
Privacy Policy Personal Data	Time Indexed Situation	
Time Indexed Participation	Task Execution	
Time Indexed Person Role	Time Indexed Person Role	

4. Finding use cases inside privacy policies

A preliminary study on the retrieval of evidences of the selected CPs inside a corpus of domain-related legal texts was performed. The study focused on a single CP, i.e. the aforementioned Privacy Policy Personal Data pattern[8]. Some evidences of it were looked for inside a small corpus of twelve privacy policies addressed to EU citizens and released after the entry into force of the GDPR. The assumption underlying the experiment is that, if an ODP should represent a recurrent ontology design problem, then evidences of this recurrence could be retrieved in the texts belonging to the domain of interest modelled by the pattern.

To verify this assumption, the text of each privacy policy was manually segmented identifying in it the paragraphs whose content was related to the semantic areas represented in the pattern. As not all the semantic areas that are relevant in a privacy policy are represented by the CP (e.g., it does not model the data subject's rights), only the paragraphs relevant for the pattern were selected. In particular, the semantic areas that were identified in it are: *(i)* types of personal data collected by the company and provided by the data subject, *(ii)* types of personal data collected by the company and provided by third parties, *(iii)* type of processing performed on personal data, *(iv)* third parties the personal data are shared with, *(v)* personal data retention period, *(vi)* lawful basis for processing. The paragraphs of the twelve privacy policies were then grouped according to the semantic area they refer to.

To automatically discover evidences of the selected CP, the ClausIE tool was applied on the paragraphs collected for each semantic area. The extracted triples were then filtered, considering those labelled by ClausIE with the label SVO, i.e. triples containing a subject (S), a verb (V) and an object (O). Finally, those triples were ordered according to the frequency they appear in the paragraphs belonging to the same semantic area. Table 3 shows the top-5 most frequent triples for each identified semantic area.

The obtained triples showed promising results for all the semantic areas. Triples that could be considered as *markers* of the presence of a relevant information to be mapped on some class of the pattern were extracted with high frequency. For instance, considering the table referring to the semantic area *(i)* (i.e., types of personal data collected from the data subject) the high frequency of the triple <we, collect, information> in the corresponding privacy policies paragraphs could be considered as an evidence of the presence in a sentence of a list of types of personal data that the company collects. Indeed, the

Table 3. Most frequent triples extracted by ClausIE and related to the six semantic area listed in Section 4. Triples in bold are the most relevant for the corresponding semantic area.

triples for semantic area *(i)*	**freq.**	**triples for semantic area *(ii)***	**freq.**
<we, collect, information>	**87**	**<we, receive, information>**	**42**
<your, “has”, information>	42	<we, collect, information>	30
<we, collect, data>	**31**	<our, “has”, games>	24
<your, “has”, device>	29	<your, “has”, information>	23
<our, “has”, website>	28	<we, collect, data>	23
triples for semantic area *(iii)*	**freq.**	**triples for semantic area *(iv)***	**freq.**
<your, “has”, information>	83	<your, “has”, information>	78
<we, use, information>	**58**	**<we, share, information>**	**76**
<your, “has”, data>	36	<your, “has”, data>	44
<our, “has”, information>	30	<your, “has”, name>	31
<your, “has”, consent>	29	**<we, share, data>**	**30**
triples for semantic area *(v)*	**freq.**	**triples for semantic area *(vi)***	**freq.**
<your, “has”, information>	41	<your, “has”, information>	25
<we, retain, information>	**27**	**<your, “has”, consent>**	**19**
<our, “has”, information>	19	**<we, process, information>**	**8**
<your, “has”, account>	16	<your, “has”, data>	7
<we, share, information>	14	<our, “has”, right>	6

privacy policies usually contain sentences like *we collect information that identifies your mobile device*. For this sentence, ClausIE extracts the following triples: <we, collect, information> and <your, “has”, device>, where the second triple is automatically inferred when the verb *to have* is preceded by a personal adjective. Thus, by analysing the frequency of each triple as well as its co-occurrence with other related triples, it could be possible to evaluate which are the concepts and the properties that a CP models and that can be retrieved inside a legal text belonging to the domain of interest. Considering the aforementioned example, each element of the triples could be mapped in some parts of the corresponding CP: the verb *collect* its an evidence for the *DataCollectionStep* class, the *your* adjective (intended as the “you” pronoun) corresponds to the *Agent* class and the *mobile device* noun could be mapped in the *PersonalData* class. Similar mappings could be identified also for the other semantic areas.

5. Conclusion and future work

This paper presents a first insight for the extraction of existing ODPs (specifically, CPs) for the data protection domain. The proposed approach uses OIE techniques to extract evidence of a CP from legal texts, aiming to achieve a fine granularity in the extraction of information. A first experiment tested the retrieval of evidences of a CP inside a small corpus of privacy policies. The next challenges to be addressed will concern the exploitation of the N-ary relations extracted by ClausIE in order to improve the retrieval of evidence of the CPs inside the text. Moreover, the evaluation of the types of legal documents where the evidence of a pattern could be looked for will be crucial for the success of the experiments.

References

[1] Harshvardhan J Pandit and Dave Lewis. “Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies”. In: *PrivOn@ ISWC*. 2017.

[2] Harshvardhan J Pandit et al. “GDPRtEXT-GDPR as a linked data resource”. In: *European Semantic Web Conference*. Springer. 2018, pp. 481–495.

[3] Monica Palmirani and Guido Governatori. “Modelling Legal Knowledge for GDPR Compliance Checking”. In: *JURIX*. 2018, pp. 101–110.

[4] Christian Bizer, Tom Heath, and Tim Berners-Lee. “Linked data: The story so far”. In: *Semantic services, interoperability and web applications: emerging concepts*. IGI Global. 2011, pp. 205–227.

[5] Aldo Gangemi and Valentina Presutti. “Ontology design patterns”. In: *Handbook on ontologies*. Springer. 2009, pp. 221–243.

[6] Cesare Bartolini, Robert Muthuri, and Cristiana Santos. “Using ontologies to model data protection requirements in workflows” . In: *JSAI International Symposium on Artificial Intelligence*. Springer. 2015, pp. 233–248.

[7] Monica Palmirani et al. “PrOnto: Privacy Ontology for Legal Reasoning”. In: *International Conference on Electronic Government and the Information Systems Perspective*. Springer. 2018, pp. 139–152.

[8] Valentina Presutti et al. “D2. 5.1: A Library of Ontology Design Patterns: reusable solutions for collaborative design of networked ontologies. NeOn Project Deliverable”. 2018.

[9] Michele Banko et al. “Open information extraction from the web”. In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. IJCAI’07. Morgan Kaufmann Publishers Inc. 2007, pp. 2670–2676.

[10] Anthony Fader, Stephen Soderland, and Oren Etzioni. “Identifying Relations for Open Information Extraction”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP ’11. Association for Computational Linguistics. 2011, pp. 1535–1545.

[11] Claudio Delli Bovi, Luca Telesca, and Roberto Navigli. “Large-scale information extraction from textual definitions through deep syntactic and semantic analysis”. In: *Transactions of the Association for Computational Linguistics*. Vol. 3. 2015, pp. 529–543.

[12] Alan Akbik and Alexander Löser. “Kraken: N-ary facts in open information extraction”. In: *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. Association for Computational Linguistics. 2012, pp.52–56.

[13] Luciano Del Corro and Rainer Gemulla. “Clausie: clause-based open information extraction”. In: *Proceedings of the 22nd international conference on World Wide Web*. ACM. 2013, pp. 355–366.

[14] Aldo Gangemi. “Ontology design patterns for semantic web content”. In: *International semantic web conference*. Springer. 2005, pp. 262–276.

[15] Valentina Presutti and Aldo Gangemi. “Content ontology design patterns as practical building blocks for web ontologies”. In: *International Conference on Conceptual Modeling*. Springer. 2008, pp. 128–141.

[16] Cristiana Santos et al. “Complaint Ontology Pattern-COP”. In: *Advances in Ontology Design and Patterns*. Vol. 32. IOS Press. 2017, pp.69–83.

[17] Harshvardhan J Pandit, Declan O’Sullivan, and Dave Lewis. “An Ontology Design Pattern for Describing Personal Data in Privacy Policies”. In: *WOP@ ISWC*. 2018, pp. 29–39.