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

doi:10.3233/FAIA190326

PrOnto Ontology Refinement Through Open Knowledge Extraction

Monica PALMIRANI[a1], Giorgia BINCOLETTO[a], Valentina LEONE[b], Salvatore SAPIENZA[a] and Francesco SOVRANO[c]
[a] *CIRSFID, University of Bologna*
[b] *Computer Science Department, University of Turin*
[c] *DISI, University of Bologna*

Abstract. This paper presents a refinement of PrOnto ontology using a validation test based on legal experts' annotation of privacy policies combined with an Open Knowledge Extraction algorithm. Three iterations were performed, and a final test using new privacy policies. The results are 75% of detection of concepts and relationships in the policy texts and an increase of 29% in the accuracy using the new refined version of PrOnto enriched with SKOSXL lexicon terms and definitions.

Keywords. legal ontology, GDPR, OKE, refinement.

1. Introduction

We have already published several papers about PrOnto ontology [17][18][19][22] that aims to model the concepts and their relationships presented in the GDPR (General Data Protection Regulation EU 2016/679). This article intends to present a validation process of PrOnto ontology using a bottom-up approach, starting from the language adopted in real examples of Privacy Policies. The research investigates: i) if the existing PrOnto classes are sufficiently exhaustive to support NLP tools in detecting GDPR concepts directly from Privacy Policies; ii) if some classes are missing with respect to the pragmatic language forms; iii) if some frequent terminology could be added to the conceptualisation modelling using e.g., SKOSXL; iv) whether it is possible to create a ML tool that is capable of detecting GDPR concepts in the Privacy Policies. The paper first presents the used methodology; secondly, it presents the legal analysis of the Privacy Policies chosen for the validation and the related mapping of the linguistic terminology in the PrOnto classes; then, the work introduces the ML technique applied to detect the PrOnto concepts from the other Privacy Policies and its results; finally, the conclusion shows the refinements made to the PrOnto ontology thanks to the validation with the Privacy Policies.

2. Methodology

PrOnto was developed through an interdisciplinary approach called MeLOn (Methodology for building Legal Ontology) and it is explicitly designed in order to minimise the difficulties encountered by the legal operators during the definition of a

[1] E-mail: {monica.palmirani, salvatore.sapienza2, giorgia.bincoletto2, francesco.sovrano} @unibo.it, leone@di.unito.it

legal ontology. MeLOn applies a top-down methodology on legal sources. It is based on reusing ontology patterns [12] and the results are evaluated using foundational ontology (e.g., DOLCE [8]) and OntoClean [11] method. The validation is made by an interdisciplinary group (engineers, lawyers, linguists, logicians and ontologists) that integrates the contributions of different disciplines. The methodology is based on the following pillars [1][3]: (i) two legal experts selected ten privacy policies from US-based companies providing products and services to European citizens; (ii) the privacy policies were analyzed using the comparative legal method to discover the frequent concepts mentioned in the texts; (iii) selected portions of text were mapped into the PrOnto ontology with also different linguistic variations; (iv) computer science team developed Open Knowledge Extraction technique starting from the GDPR lexicon, PrOnto ontology and the literal form variants (point 3); (v) results were validated by the legal team that returned them to the technical team; (vi) the steps from (ii) to (v) were iterated three times to refine the ontology and the software model; (vii) finally, new privacy policies were selected by the legal experts[2] in order to evaluate the effectiveness of the refined algorithm and ontology.

3. Legal Analysis of the Privacy Policies

We have selected ten Privacy Policies[3] from an equal number of companies in the sector of sale of goods, supply of services and sharing economy. We chose these companies due to their international dimension, their relevance in their market sectors and the diversity of data processing techniques, with European target. We distinguished between the legal strict terminologies (e.g., data subject) to the communicative language (e.g., customer or user). The legal experts have manually reviewed the Privacy Policies to discover the concepts of legal relevance for data protection domain (provisions, legal doctrine, WP29 and case law) that are remarkably recurrent in the text. The interpretation has also kept into account the existing version of PrOnto ontology, in particular to identify the different terms that express the same concept recognised through a legal analysis at an equal level of abstraction. These terms have been analysed, compared and eventually included in the PrOnto ontology, using techniques like SKOSXL for adding the different linguistic forms (e.g., `skosxl:leteralForm`). This extension of PrOnto definitely improves the capacity of the OKE tools to detect the correct fragment of text and to isolate the legal concept as well as populating the PrOnto ontology. We also noted that the Privacy Policies tend to use the ordinary, everyday language for reasons of transparency and comprehensibility of the texts. Despite the advantage for the costumer/user, the analysis underlined that certain terminologies are not accurate from a legal perspective. For instance, the expression "*giving permission*" is a communicative substitute of "*giving consent*" and "*obtain consent*". Some terminologies are misused because the ordinary language in the policy does not reflect the legal sense e.g., "*anonymous data*" (Recital 26 GDPR) is not in the scope of the Regulation and it is misled with "*anonymized data*". We found terminology coming from computer science like "*to hash*", "*log files*", "*use encryption*" convey a technical meaning that is not classified in the GDPR, which is drafted in a technically neutral way.

[2] Rover, Parkclick, Springer, Zalando, Louis Vuitton, Burger King, Microsoft-Skype, Lufthansa, Booking, Zurich Insurance.

[3] Amazon, Dell, McDonald, Nike, American Airlines, TripAdvisor, Hertz, Allianz U.S. AirBnB, Uber.

4. PrOnto Manual Enhancing

Following this analysis, we have mapped the synthesis of the different lexicon expressions with the PrOnto classes. This step allowed to detect some missing modules that are described below. Under the GDPR, personal data processing (Art. 4.1(2)) is lawful only if motivated by a purpose that must be legitimated by a legal basis (Art. 6 GDPR). Therefore, a lawfulness status was thus added as a Boolean data property of the `PersonalDataProcessing` class. However, from the validation using Privacy Policies, it is extremely important to elicit the Legal Basis because several other implications (rights, obligations, actions) depends to the kind of legal basis (e.g., Art. 22). For this reason, we have modelled new module (Fig. 1 new classes are in orange).

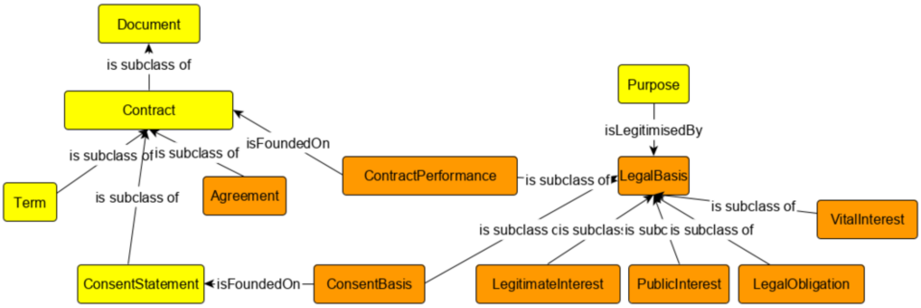

Figure 1 – Legal Basis Module

Archiving and Services are encountered frequently in the Privacy Policies and they are added to the Purpose Module, with also a specific kind of service (`InformationSocietyService`) relevant for the child privacy (Art. 8 GDPR). The Privacy Policies underlined some obligations, and related rights, like the `ObligationToProvideHumanIntervention` connected with `RightToHaveHumanIntervention` and related with `AutomaticDecisionMaking` that is an action added to the Action module.

5. Open Information Extraction for PrOnto

We built a software for detecting GDPR concepts from Privacy Policies taking inspiration from the PrOnto ontology and using a tool conceptually based on ClausIE [6]. ClausIE is a clause-based approach to Open Information Extraction, which extracts relations and their arguments from natural language text. Open Information Extraction (Open IE) builds information graphs representing natural language text in the form of SVO (Subject, Verb, Object) triples (slightly different from RDF). This method was used in other relevant works in the past and several problems arise: (i) linguistic variants of the same legal concept inside the agreement/contract text are numerous and they include some overlappings of meaning; (ii) while legislative text uses rhetoric sentences, policy text is usually simpler and uses common language to be more understandable; (iii) occasionally, legal provisions are written in passive form in order to emphasize prescriptiveness when addressing the command; (iv) legal text has normative references that affect the knowledge extraction; (v) legal concepts change over time; (vi) frequency is not a good indicator of relevance. The main difference between many classical Open IE techniques and ClausIE is that the latter makes use of the grammatical dependencies extracted through an automatic dependency parser, to identify the SVO triples. ClausIE is able to identify SVO triples, but we need also to correctly associate them to ontology terms and their literal variants provided by the legal expert team. Let the GDPR and the Privacy Policies be our corpus C. In order to perform the automatic text annotation of our corpus with PrOnto concepts, we follow these steps: 1. we identify a list of all the terms (subjects, objects-classes; verbs-properties) in C, by using a simple variant of ClauseIE; 2. we use PrOnto labels of classes and properties, with additional mapping of linguistic and lexicon variants; 3. we map every possible class/property in C to its closest class/property in PrOnto, using a previous project[4]. This algorithm exploits pre-trained linguistic deep models in order to easily compute a similarity score between two terms.

6. PrOnto Refinement Using OKE

From the Privacy Policies linguistic analysis with OKE, it emerges that some inputs produced important enhancements in PrOnto ontology. **New Child Class**: in the Privacy Policies is frequently mentioned “*child*” that is a particular “data subject” missing in the PrOnto ontology. Initially, we intended to use rules to define child concept because the definition changes for each jurisdiction according to the local implementation of the EU Regulation. However, in light of the important rights and obligations defined in the GDPR for the minors, we decided to include a new class in the `Role` module as subclass of `DataSubject`. `Child` class is related with `ParentalResponsabilityHolder`. **New AnonymisedData Class**: from the Privacy Policies linguistic analysis emerges that “*Anonymised Data*” and “*Anonymous data*” (Recital 26 GDPR)[5] are often misled. The pragmatic language attempts to simplify the legal terminology creating mistake in the conceptualization of those two classes. To stress this distinction, we modelled the relationship `PersonalData isTransformedIn AnonymisedData`.

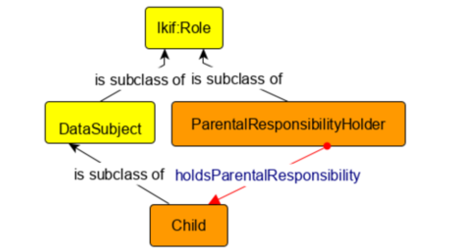

Figure 2: Child class.

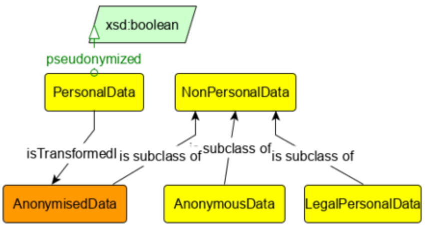

Figure 3: AnonymisedData class.

The best manner to detect an *action* is through verbs. However, within OWL ontology, verbs play the role of predicates that connect domain and range (relationships not classes). For this reason, the legal team modifies the action’s classes with the “ing” form according also other scholars [10]. **New Actions** are detected like `Collecting` and `Profiling`. The legal analysis collocates the `Profiling` class as subclass of `AutomatedDecisionMaking` following Art. 22 and the Recital 71. In this case, the OKE provides a very good input to the legal experts that provided an improvement of the legal ontology by relying on their legal analysis. **Lexicon Forms**: it is important to connect the legal concepts to lexicon form variants. We use SKOS and SKOSXL that is

[4] https://gitlab.com/CIRSFID/un-challange-2019.

[5] COM (2019) 250 final anonymised “data which were initially personal data, but were later made anonymous.”. Recital 26 GDPR “6. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.”

a canonical method for connecting OWL and linguistic variants, using `skosxl:literalform`. In this manner, we link PrOnto Core Ontology with other existing lexicon-controlled vocabulary[6].

7. Related Work

Several ontologies model privacy domain. Some of them are oriented to the linguistic tools e.g., UsablePrivacy and PrivOnto [15] to define glossary starting from the bottom-up annotation of the privacy policies (crowdsourcing annotation). GDPRtEXT [20] lists concepts present in the GDPR text without claiming to model norms and legal axioms. GDPRov describes the provenance of the consent and data lifecycle in Linked Open Data [21]. GConsent models the consent action, statement and actors. The SPECIAL Project develops tools for checking compliance in privacy domain. ODRL provides predicates and classes for managing obligations, permission, prohibitions, but not deontic logic operators (e.g., penalty). LegalRuleML [16] ontology was included inside of PrOnto. EUROVOC and IATE are some examples of linguistic ontologies released by the European Union to semantically structure the terminology of documents of the EU institutions [23]. Those resources do not clarify the distinction between legal concepts and their instances and additional knowledge is necessary on legal theory, legal doctrine and legal sociology [7]. Several models propose interfaces between high-level ontological concepts and their low-level, context-dependent lexicalisations [14]. SKOSXL[5] and OntoLex [4] are included in this version of PrOnto for combining ontology and linguistic literal forms, in support to NLP and search engine. Open IE is capable to extract information graphs from natural language. Examples of Open IE tools are ClausIE [6], OpenCeres [13] and Inter-Clause Open IE [1]. Open Knowledge Extraction (Open KE) builds over Open IE to align the identified subject, predicates and objects (SVOs) to pre-defined ontologies. FRED [9] uses different NLP techniques for processing text and for extracting a raw ontology based on VerbNet situations. The challenge of Open KE is that the SVOs alignment requires to understand the meaning of ambiguous and context-dependent terms. Our algorithm tackles the Open KE problem by exploiting pre-trained linguistic deep models to map information to knowledge.

8. Conclusions and Future Work

We have validated the PrOnto ontology with a sample of Privacy Policies and with a legal analysis following the MeLOn methodology, in order to manually check the completeness of the classes and relationships for representing the main content of the policies texts. This exercise detected some new classes in the PrOnto ontology (e.g., Legal Basis). The legal team detected some inconsistency in the terminologies between the legislative text and the pragmatic language. This produced a map of lexicon variants, then modelled using SKOSXL. PrOnto and these extensions fill up an OKE algorithm to detect concepts in the Privacy Policies. The method was iterated three times and at the end we obtained an increase of 29% in the detection of the concepts respect the first interaction that record an increase of 19%. We are capable to detect the 75% of the concept in the new privacy policies using the new version of PrOnto enriched with SKOSXL terms. This method is also relevant to annotate legal texts with PrOnto and so to create RDF triples for supporting applications (e.g., search engine, legal reasoning)[7].

[6] https://www.w3.org/ns/dpv#data-controller.

[7] https://gitlab.com/palmirani/pronto.

Acknowledgements. This work was partially supported by the European Union's Horizon 2020 research and innovation programme under the MSCA grant agreement No 690974 “MIREL: MIning and REasoning with Legal texts”.

References

[1] G. Angeli, M.J.J. Premkumar, C.D. Manning. Leveraging linguistic structure for open domain information extraction. In ACL-IJCNLP **1** (2017), 344–354.

[2] K.D. Ashley, *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge Univ Press, Cambridge New York, 2017.

[3] J. Bandeira, I.I. Bittencourt, P Espinheira, S. Isotani. FOCA: A Methodology for Ontology Evaluation. ArXiv preprint arXiv:1612.00353 (2016).

[4] J. Bosque-Gil, J. Gracia, E. Montiel-Ponsoda. Towards a module for lexicography in OntoLex. In Proc. of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets at LDK 2017, Galway. CEUR-WS **1899** (2017), 74–84.

[5] T. Declerck, K. Egorova, E. Schnur. An Integrated Formal Representation for Terminological and Lexical Data included in Classification Schemes. In Proc. of the LREC-2018 (2018).

[6] L. Del Corro, R. Gemulla. Clausie: clause-based open information extraction. In Proc. of the 22nd intern. conference on World Wide Web. ACM (2013), 355-366.

[7] M. Fernández-Barrera, G. Sartor. The legal theory perspective: doctrinal conceptual systems vs. computational ontologies. In Approaches to Legal Ontologies. Springer, Dordrecht (2011), 15–47.

[8] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, L. Schneider. Sweetening Ontologies with DOLCE. In Inter. Conf. on Knowledge Engineering and Knowledge Management Springer, Springer, Berlin, Heidelberg, (2002) 166–181.

[9] A. Gangemi, A. Presutti, V. Reforgiato, D. Recupero, A.G. Nuzzolese, F. Draicchio, M. Mongiovì. Semantic web machine reading with FRED. *Semantic Web*, **8** (2017), 873–893.

[10] A. Gangemi, S. Peroni, D. Shotton, F. Vitali. The Publishing Workflow Ontology (PWO). *Semantic Web* **8** (2017), 703–718.

[11] N. Guarino, C.A. Welty. An Overview of OntoClean. In Handbook on ontologies (2004), 151–171.

[12] P. Hitzler, A. Gangemi, K. Janowicz, A. Krisnadhi (Eds.). *Ontology engineering with ontology design patterns: foundations and applications, Studies on the semantic web*. IOS Press, Amsterdam. 2016

[13] C. Lockard, P. Shiralkar, X. L. Dong. OpenCeres: When Open Information Extraction Meets the Semi-Structured Web. In *NAACL-HLT 2019* **1** (2019), 3047–3056.

[14] J. McCrae, D. Spohr, P. Cimiano. Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In Proc. ESWC 2011. *LNCS* **6643** (2011), 245–259.

[15] A. Oltramari, D. Piraviperumal, F. Schaub, S. Wilson, S. Cherivirala, T.B. Norton, N.C. Russell, P. Story, J. Reidenberg, N. Sadeh. Privonto: A semantic framework for the analysis of privacy policies. *Semantic Web* **9** (2018), 185–203.

[16] M. Palmirani, G. Governatori. Modelling Legal Knowledge for GDPR Compliance Checking. In Proc. Jurix 2018. *Frontiers in Artificial Intelligence and Applications* **313** (2018), 101–110.

[17] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, L. Robaldo, 2018. PrOnto: Privacy Ontology for Legal Reasoning. In Proc. EGOVIS2018, September 3-5. *LNCS* **11032** (2018), 139–152.

[18] M. Palmirani, M. Martoni, A. Rossi. C. Bartolini, L. Robaldo. Legal Ontology for Modelling GDPR Concepts and Norms. In Proc. JURIX 2018. *Frontiers in Artificial Intelligence and Applications* **313** (2018), 91–100.

[19] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, L. Robaldo. PrOnto: Privacy Ontology for Legal Compliance. In Proc. ECDG 2018, ACPI Reading UK (2018), 142–151.

[20] H.J. Pandit, K. Fatema, D. O'Sullivan, D. Lewis. GDPRtEXT - GDPR as a Linked Data Resource. In Proc. ESWC 2018. *LNCS*, **10843** (2018), 481–495.

[21] H.J. Pandit, D. Lewis. Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies. In Proc. of the 5th Workshop PrivOn2017 co-located with ISWC 2017 (2017)

[22] A. Rossi, M. Palmirani, 2019. DaPIS: an Ontology-Based Data Protection Icon Set. *Frontiers in Artificial Intelligence and Applications* **317** (2018), 181–195.

[23] C. Roussey, F. Pinet, M.A. Kang, O. Corcho. An introduction to ontologies and ontology engineering. *Ontologies in Urban development projects* **1** (2011) 9–38.