Legal Knowledge and Information Systems
S. Villata et al. (Eds.)

doi:10.3233/FAIA200851

The Role of Vocabulary Mediation to Discover and Represent Relevant Information in Privacy Policies

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

Abstract. To date, the effort made by existing vocabularies to provide a shared representation of the data protection domain is not fully exploited. Different natural language processing (NLP) techniques have been applied to the text of privacy policies without, however, taking advantage of existing vocabularies to provide those documents with a shared semantic superstructure. In this paper we show how a recently released domain-specific vocabulary, i.e. the Data Privacy Vocabulary (DPV), can be used to discover, in privacy policies, the information that is relevant with respect to the concepts modelled in the vocabulary itself. We also provide a machine-readable representation of this information to bridge the unstructured textual information to the formal taxonomy modelled in it. This is the first approach to the automatic processing of privacy policies that relies on the DPV, fuelling further investigation on the applicability of existing semantic resources to promote the reuse of information and the interoperability between systems in the data protection domain.

Keywords. legal vocabularies, ontology population, text similarity, data protection

1. Introduction

In the European Union (EU), the entry into force of the General Data Protection Regulation (GDPR) [1] has brought the domain of data protection to the forefront, encouraging the research in knowledge representation and natural language processing (NLP), among the other fields. On the one hand, several ontologies and vocabularies adopted Semantic Web standards to provide a formal representation of the data protection framework set by the Regulation. On the other hand, different NLP approaches have been applied to the text of privacy policies to address classification tasks that assign one or more labels to the paragraphs of a privacy policy, according to its content.

These two lines of research do not seem to pursue a common goal. The labels used in the classification tasks are not organised in a semantic structure and the outcomes of these tasks are hardly applicable outside the context of the project for which they were implemented. Consequently, the full potential of Semantic Web oriented vocabularies is not exploited to provide the text of privacy policies with a shared semantic superstructure and their effort to promote interoperability between systems on the Web is lost.

[1] Corresponding Author. E-mail: valentina.leone@unito.it

Among the most recent semantic resources that were proposed to model the data protection domain, the Data Privacy Vocabulary (DPV) [2] organises its concepts in a lightweight taxonomic structure. This vocabulary has drawn the attention of many projects that declared their interest in its adoption [3,4,5]. However, to the best of our knowledge, no effort has been yet made to automatically extract, from privacy policies, the information that is relevant with respect to the concepts modelled by this vocabulary.

In this paper we present a first approach that is driven by the concepts modelled in the DPV to automatically discover the relevant information in privacy policies. The proposed method integrates the knowledge represented in the DPV with the information modelled in BabelNet[2] [6], i.e. a general-purpose vocabulary that provides a semantic network of concepts linked through lexical and semantic relationships. The outcome of the method is provided in a machine-readable format that bridges the gap between the unstructured text of a privacy policy and the formal taxonomy of concepts provided by the DPV. The paper in structured as follows: Section 2 presents some related work and Section 3 describes the resources that we adopted in our experiments. Section 4 explains the steps that were implemented to map the information in the privacy policies on the concepts of the DPV, while Section 5 explains how the proposed methodology was evaluated. Section 6 describes the machine-readable representation that has been provided for the results and Section 7 ends the paper with some final remarks.

2. Related Work

Many works in the data protection field applied NLP techniques to the text of privacy policies for labelling their paragraphs according to the information they express. Polisis [7] relies on domain-specific word embeddings and a hierarchy of neural networks to classify the paragraphs of the policies in the OPP-115 corpus [8]. The approach described in Polisis is then refined and improved in [9]. Supervised machine learning models are also used in PrivacyGuide [10] to highlight the risk level associated to some privacy aspects described in privacy policies. An unsupervised learning technique is adopted in [11] to extract the topics emerging from a corpus of more that 4K privacy policies, then comparing those topics with the information represented by the labels provided in the OPP-115 corpus. KniGHT [12] exploits techniques of semantic text matching for mapping the sentences of a privacy policy on the most related articles of the GDPR.

Other projects focused on representing the information originating from different textual sources in the data protection field into structured machine-readable representations. The Lynx[3] project aims, in one of its use cases, to create a knowledge graph for the data protection field interlinking domain-related legal texts and providing algorithms able to automatically enlarge the knowledge base when new relevant documents are issued [13]. The SPECIAL[4] project [3] focused on the development of machine-readable policy languages and the DPV was proposed in the context of this project. The MIREL[5] project included the PrOnto ontology among its outcomes, proposing a technique based on Open Information Extraction to map the information extracted from privacy policies on the classes modelled by the ontology [14].

[2] https://babelnet.org/
[3] http://www.lynx-project.eu/
[4] https://www.specialprivacy.eu/
[5] https://mirelproject.eu/index.html

Table 1. The modules in the DPV and the labels in the OPP-115 corpus for paragraph-level annotations.

DPV Modules	Personal Data Category; Processing; Purpose; Legal Basis; Data Controller; Recipient; Data Subject; Technical Organisational Measures
OPP-115 Labels	First Party Collection/Use; Third Party Sharing/Collection; User Choice/Control; User Access, Edit & Deletion; Data Retention; Data Security; Policy Change; Do Not Track; International & Specific Audiences; Other

3. Scope and Limitations of the Adopted Resources

In our experiments, the DPV was jointly used with a corpus of privacy policies, named OPP-115 corpus[6] [8]. The different nature and scope of these resources required us to put some constraints on their use.

The Data Privacy Vocabulary[7] [2] was first released in July 2019. Through a formal representation that relies on RDF and OWL, it aims to provide a basic vocabulary of terms related to the data protection domain framed by the GDPR. The DPV is made of several modules, that provide a taxonomy of terms related to different aspects involved in the personal data handling. Those modules are listed in the first row of Table 1.

The OPP-115 corpus includes 115 privacy policies that were manually labelled with a two layered annotation made at paragraphs and text spans level. The paragraphs are associated with labels representing ten different data practices, listed in the second row of Table 1. Within an annotate paragraph, text spans are labelled with attribute-value pairs that are specific for a given data practice and that can assume a limited set of values. The OPP-115 corpus collects privacy policies that were issued by US-companies some years before the entry into force of the GDPR. Therefore, some concepts modelled in the DPV can not be expected to be mentioned in the corpus.

To take into account of the different scope of the resources, the extraction of relevant information from the text of the privacy policies is limited to the concepts in the *Personal Data Category* and *Purpose* modules in the DPV. Similarly, we only considered the paragraphs of the privacy policies that were assigned to the *First Party Collection/Use* label in the corpus, as we expect this information to be more likely to be found within them. From here on, even if no further specified, we assume that the implemented method and its evaluation were applied taking into account the aforementioned constraints in the joint adoption of the two resources.

4. Method

The method described in this Section is composed of three sequential steps, where the output of one step becomes the input of the following one. The first step creates broad mappings between some parts of the text in the privacy policies and the modules of the DPV. The second step tries to refine these mappings selecting, from the modules in the DPV, some classes that could be suitable for the refinement. The last step chooses, from the set of suitable classes, the one that will yield the needed refinement.

[6] https://www.usableprivacy.org/data

[7] https://dpvcg.github.io/dpv/

4.1. Broad Mappings of Text Chunks on the DPV Modules

To discover the parts of the text in privacy policies that are relevant with respect to the DPV, the first step of the method leverages the distinctiveness of the terminology that characterises each module of the vocabulary. This evidence was found collecting and ordering, by decreasing frequency, the terms used to name the classes in each module and to provide the description of their meaning in natural language (through the RDF property `dct:description`). As Table 2 shows, the collected terms are in most of the cases exclusive for each module and only few words overlap. Thus, the nouns in each list can be considered as *descriptors* for the type of information that each module of the DPV represents. Moreover, for each descriptor, we also considered its synonyms, that were automatically retrieved from BabelNet.

The discovery of relevant parts of the text in the privacy policies relied on these descriptors. For each sentence, the noun chunks (i.e. the nominal phrases) were extracted using the available libraries of the SpaCy dependency parser[8] and the chunks roots (i.e. the words connecting the noun chunks to the rest of the parsed sentence) were used to perform the mappings. When the root of a chunk matched a descriptor, the chunk was mapped on the corresponding module. In case of a match with a descriptor that appears in both the modules, the chunk was assigned to the module where the descriptor has the highest frequency. In case of a tie, the chunk was preliminarily assigned to both modules. The chunks whose roots did not find a match with a descriptor were considered not relevant in establishing a match with the DPV. Two examples of the mappings performed in this step are shown below. The module assigned to each chunk is indicated in a square box and the roots of the chunks, used to determine the mappings, are underlined.

$$\boxed{\text{Purpose}} \quad \textit{customer service } \underset{\text{root}}{\underline{\textit{purpose}}}$$

$$\boxed{\text{Personal Data Category}} \quad \textit{mobile device unique id } \underset{\text{root}}{\underline{\textit{number}}}$$

4.2. Detection of Candidate Classes for the Refinement of the Broad Mappings

Given the coarse assignments of noun chunks to one or two modules in the DPV, the second step focused on the refinement of these assignments identifying a set of more specific candidate classes in the taxonomies of the modules. Given a text chunk, a first control checks if the name of a class in the DPV, or one of its synonyms retrieved with BabelNet, matches the chunk or appears as a sub-string in it. If this is the case, the set of

[8] https://spacy.io/

Table 2. The six most frequent words used to name and describe the classes in the *Purpose* and *Personal Data Category* modules of the DPV. The number next to each noun represents the frequency of the noun in the module. More than six terms are present in both lists due to the tie in the frequencies.

DPV Module	**Top-6 of the frequent words**
Purpose	(service, 17), (user, 9), (product, 8), (research, 8), (optimisation, 7), (datum, 6), (activity, 6), (commercial, 6), (recommendation, 6), (interface, 4), (individual, 4), (purpose, 4)
Personal Data Category	(individual, 148), (information, 141), (history, 18), (health, 17), (personal, 17), (social, 13), (credit, 13), (datum, 13), (professional, 11)

candidate classes is made of a single element, i.e. the matching class. For instance, the fragment considered in the previous Section:

$$\boxed{\text{Purpose}} \quad \underset{\text{dpv:CustomerCare}}{\underline{\textit{customer service}}} \quad \underset{\text{root}}{\underline{\textit{purpose}}}$$

contains the sub-string *customer service* that is a synonym of the string *customer care*. In turn, the latter matches the homonym DPV class, that is considered as a candidate class to perform the refinement of the mapping with the *Purpose* module.

If no class is detected with this first check, then the lists of modules descriptors (see Section 4.1) are used to populate the list of candidate classes. Specifically, for each descriptor that matches a word in the text chunk, the class from which the descriptor was extracted is added to the list of candidate classes. If a candidate class is a leaf in the taxonomy of a module, then it is substituted by its direct superclass in order to avoid matches with too specific classes. The root of the text chunk is excluded from the search of the candidate classes, because it already contributed to the broad mappings with the DPV modules. For instance, in the following fragment:

$$\boxed{\text{Personal Data Category}} \quad \textit{mobile} \quad \underset{\text{dpv:DeviceBased}}{\underline{\textit{device}}} \quad \underset{\text{dpv:Identifying}}{\underline{\textit{unique}}} \quad \textit{id} \; \underset{\text{root}}{\underline{\textit{number}}}$$

the word *device* matches a descriptor that corresponds to the *DeviceBased* class, while the word *unique* matches a descriptor corresponding to the *UID* (i.e. user identifier) class. However, as the class is a leaf in the taxonomy of the *Personal Data Category* module, its direct superclass, i.e. *Identifying*, is added to the set of candidate classes of the chunk.

4.3. Selection of the Class for Refining the Broad Mappings

The third and last step of the method selects, among the candidate classes, the most suitable for refining the broad mapping between a text chunk and a module. Following many simple but consolidated state of the art approaches [15,16], the class is selected by computing the cosine similarity between the text chunk and its candidate classes. The vector representation of both the text chunk and its candidate classes was obtained from the pre-trained GloVe word embeddings[9] [17] that were combined according to some weights for representing the different contributions given by each word in the overall vector representation.

The vector representation for a text chunk is obtained collecting the set W_F of the embeddings for the content words in the chunk and the set W_S of the embeddings for the content words that occur in the same sentence of the text chunk. Assuming that all the words in the chunk contribute equally to its vector representation, a weight equal to 1 is assigned to each word embedding in W_F. By contrast, the weights associated to the word embeddings in W_S assume that the contribution of a word occurring in the same sentence of the chunk is equal to the frequency of that word in the sentence divided by the total number of distinct words in the sentence. The vector representation of the text chunk is computed, then, by multiplying each embedding in the set W_F and W_S by the corresponding weight and computing the mean vector resulting from the two sets.

By contrast, the vector representation for a candidate class is conceptually based on the computation of some Term Frequency-Inverse Document Frequency (TF-IDF) scores

[9] https://nlp.stanford.edu/projects/glove/. We use the 300-dimensional vectors.

Table 3. Statistics about the number of text chunks that were retrieved in the privacy policies and the number of classes of the DPV that were associated with at least a text chunk.

	Purpose	**Personal Data Category**	**Total**
Chunks (with repetitions)	852	4025	4877
Chunks (no repetitions)	224	747	971
Retrieved classes	17	85	102

associated to the words used in its description. Following the assumption that underpins the TF-IDF measure, terms that are used in one or few class descriptions should be emphasised, because they likely are more representative of a specific DPV class, while terms that are used frequently in the definitions of the classes should have less relevance. Therefore, being C the set of candidate classes for a text chunk, the TF-IDF scores for the content words used in the description of a class c in C were computed considering the frequency of these words in the description of c and the inverse document frequency of these words with respect to the definition of the other classes in C. The embeddings for the content words in the description of c were then multiplied by the corresponding TF-IDF scores and the average vector of the embeddings was computed to obtain the vector representation of c.

Finally, the cosine similarity is computed between the vectorial representations of the chunk and of each candidate class. The class that results in the highest cosine similarity value is considered as the best candidate for the refinement. The example below shows the similarity values computed for the text chunks discussed in the previous sections (the class that determined the final mapping is highlighted in bold).

$$\boxed{\text{Purpose}}\quad \underset{\textbf{dpv:CustomerCare 0.77}}{\underline{\textit{customer service}}}\quad \underset{\text{root}}{\underline{\textit{purpose}}}$$

$$\boxed{\text{Personal Data Category}}\ \textit{mobile}\quad \underset{\textbf{dpv:DeviceBased 0.76}}{\underline{\textit{device}}}\quad \underset{\text{dpv:Identifying 0.60}}{\underline{\textit{unique}}}\quad \textit{id}\quad \underset{\text{root}}{\underline{\textit{number}}}$$

5. Evaluation

5.1. Statistics about the Performed Mappings

Table 3 shows a summary of the number of text chunks that were extracted with the methodology described in Section 4. Overall, we extracted 4877 chunks that were associated to 102 classes of the DPV (out of a total of 192 classes in the two modules of interest). Each chunk occurs one or more times in the corpus of privacy policies. Omitting the repetitions, the number of unique text chunks that were retrieved is equal to 971. Among them, 128 chunks were detected because the name of a class (or one of its synonyms) matched the chunk or appeared as a sub-string in it. The remaining 843 chunks were retrieved populating the lists of candidate classes, relying on the descriptors extracted for each module (see Section 4.2).

5.2. Precision Assessment of the Performed Mappings

The evaluation of the results relied on the annotations of the privacy polices provided by the OPP-115 corpus (see Section 3 for further details).

Table 4. DPV classes with the highest and lowest number of text chunks mapped on them.

	Personal Data Category	Purpose
Most Frequent	(Device Based, 758), (Email Address, 282), (Contact, 183)	(Commercial Interest, 337), (Purpose, 266), (Security, 49)
Less Frequent	(Philosophical Belief, 1), (Disciplinary Action, 1), (Thought, 1)	(Access Control, 1), (Service Optimization, 1), (Optimisation For Consumer, 7)

To estimate the precision of the mappings extracted by our method, we created a correspondence between the values of the *Personal Information Type* attribute of the OPP-115 corpus and some of the DPV classes in the *Personal Data Category* module. Those correspondences were manually identified analysing the descriptions provided both for the attribute values in the corpus and the classes in the DPV, unravelling similarities in the type of information that they represent. Table 5 shows the mappings that we considered. In this table, the numbers between squared brackets represent the level of a class in the taxonomy of the module (we say that the most general class in the taxonomy lies at level 0, all its direct subclasses lie at level 1, and so on). Most of the correspondences were made between an attribute value and a class at the second level in the module. We found that some attribute values are very general and no meaningful correspondences were found. A similar analysis was also performed on the values of the *Purpose* attribute in the OPP-115 corpus and the classes of the homonym module in the DPV. Table 6 shows the mappings that we considered. In this case, most of the attribute values were associated with classes at level 1 in the *Purpose* module.

Based on the correspondences that we drawn, we identified three different scenarios for the evaluation. Given a text chunk f that is extracted from a sentence s in a privacy policy: *(i)* f is part of a text span in s and the attribute-value pair associated to the span matches the class of f, following the correspondences that were identified for the evaluation; *(ii)* f is part of a text span that is labelled in s, but the attribute-value pair associated to the span do not match the class associated with f; *(iii)* f does not correspond to any of the text spans that were annotated in s. We computed the number of text chunks that

Table 5. Correspondences between the values of the *Personal Information Type* attribute in the OPP-115 corpus and the classes in the *Personal Data Category* DPV module. The last row lists the attribute values that did not find a match in the module.

Attribute Values	Classes in the *Personal Data Category* Module
Financial	Financial [1]
Health	Medical Health [2]
Contact	Contact [2], Name[3]
Location	Location [2]
Demographic	Demographic [2], Physical Characteristic [2], Professional [2], Family [2]
Personal identifier	Identifying [2], Financial Account[2]
User online activities	Behavioral [2], Social Media Communication [3]
User profile	Identifying[2], Preference[2]
Social media data	Social Network [2]
IP address & device ids	Device Based [2]
Computer information	Device Based [2]
Cookies & traking elements, Survey data, Generic personal information, Other, Unspecified	

Table 6. Correspondences between the values of the *Purpose* attribute in the OPP-115 corpus and the classes in the *Purpose* DPV module. The last row lists the attribute values that did not find a match in the module.

Attribute Values	**Classes in the *Purpose* Module**
Basic service/feature	Service Provision [1]
Additional service/feature	Service Provision [1], Service Personalization [1]
Advertising	Service Personalization [1]
Marketing	Commercial Interest [1], Service Personalization [1]
Analytics/research	Research And Development [1], Service Optimization [1]
Personalisation/Customisation	Service Personalization [1]
Service Operation & Security	Security [1]
Legal Requirement, Merger/Acquisition, Other, Unspecified	

fit each of the three scenarios and we collected the results in Table 7. Some insights from the evaluation are presented in the next Section.

5.3. *Insights from the Results of the Evaluation*

The first insight that comes from the retrieved mappings concerns the coverage of the two modules of interest in the DPV with respect to the classes that were associated to some text chunks in the privacy policies (see the last row of Table 3). The number of classes that were automatically mapped on the text chunks slightly exceeds (53.1%) half of the concepts represented in the DPV modules of interest. However, it should be noticed that many concepts in the DPV are very specific and likely difficult to find in the privacy policies text. Classes like *Music* or *Accent* in the *Personal Data Category* module were not mapped on any text chunk. By contrast, chunks related to the *IP Address, Location* and *Contact* classes were frequently extracted. This intuition is reinforced by looking at Table 4, that provides an excerpt of the classes for which the highest and lowest number of text chunks (considering repetitions) was found.

Concerning the evaluation technique explained in Section 5.2, we noticed that most of the labels mismatches occurred because the text spans in the corpus were associated to general labels (like *Other*). We therefore believe that, in this case, our vocabulary-driven approach could provide an advantage over the manual annotation proposed in the corpus, suggesting more precise labels for the text spans. By contrast, the scenario in which a text chunk, that was automatically extracted by our method, was not annotated in the corpus needs further investigations for evaluating to what extent the lack of an annotation indicates an incorrect automatic mapping or is rather a corpus fault. However, we believe that the results obtained with this first evaluation approach, based solely on the labels provided by the corpus, have provided promising insights that encourage the refinement of the evaluation approach, that could involve manual expert evaluation.

Table 7. Results of the evaluation that is based on the manual drawing of the correspondences between attribute values in the OPP-115 corpus and the classes in the DPV according to the three different scenarios discussed in Section 5.2. Percentages are computed with respect to the total number of noun chunks extracted for the corresponding module.

	Purpose	**Personal Data Category**	**Total**
Match	114 (13.4%)	1351 (33.6%)	1465 (30.0%)
Mismatch	296 (34.7%)	858 (21.3%)	1154 (23.7%)
No annotation	442 (51.9%)	1816 (45,1%)	2258 (46.3%)

6. Semantic Web Oriented Representation of the Results

We propose a machine-readable representation of the mappings that were automatically extracted by our method. The understanding that we would like to provide about the proposed mappings is that of *semantic domains* that are identified by the concepts of the DPV, and *domain elements* that correspond to the text chunks and that are related to the semantic domains. A standardised modelling solution to this intuition is provided by the *Collection* Ontology Design Pattern (ODP)[10], that can be used to represent the membership to a domain, not to be intended in the sharp sense defined in the set theory (as specified by the documentation provided for the ODP).

We used the RDF syntax to formalise the mappings extracted from the privacy policies by using the representational model provided by this ODP. For each DPV class that was associated to a text chunk in a privacy policy, a new class representing a related semantic domain was introduced. The text chunks were then associated to their semantic domains with the property `isMemberOf`, introduced by the ODP. The properties `skos:label` and `skos:example` were used to associate to the chunks their natural language strings and the sentences of the privacy policy from which they were extracted, as shown in the example below.

```
:DemographicDomain rdf:type dpv:Demographic, owl:Thing.

:DemographicAnalysisConcept rdf:type skos:Concept, owl:Thing;
  odp:isMemberOf :DemographicDomain;
  rdfs:label "demographic analysis"@en;
  skos:example "Perform statistical, demographic, and marketing
  analyses of users of the Sites and their purchasing patterns"@en.
```

This example shows the advantage of the proposed representation: an unstructured delivery of the results could erroneously suggest that, intuitively, if the concept *Demographic* contributed to the identification of the text chunk *demographic analysis*, then there is a close match between their meanings. By contrast, the representation of a semantic domain related to the *Demographic* concept and the association of the text chunk to this domain provides a new perspective on the proposed mapping. Indeed, *demographic analysis* and *demographic personal data* are different in their meaning, but it is likely that a demographic analysis will involve the processing of demographic personal data, thus legitimating a mapping of the text chunk with the corresponding domain.

7. Conclusion and Future Work

In this paper we presented the first approach that exploits a recently-released vocabulary for the data protection domain to discover the relevant information in the text of privacy policies. Moreover, we presented a machine-readable representation of the results, based on RDF and a standardised ontological solution. The obtained results show that NLP approaches in the data protection domain can benefit from existing semantic resources, to share information and promote interoperability between systems. We plan to continue the work on the refinement of the proposed approach applying it to a corpus of GDPR-compliant privacy policies. This would make it possible to overcome some of the aforementioned limitations of the OPP-115 corpus and to extend the applicability of the DPV to other modules.

[10] http://ontologydesignpatterns.org/wiki/Submissions:Collection

References

[1] 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. p. 1-88 (May 2016).

[2] Harshvardhan JP, Polleres A, Bos B et al. Creating a vocabulary for data privacy: the first-year report of data privacy vocabularies and controls community group (DPVCG). 2019. In: Panetto H, Debruyne C, Hepp M, et al., editors. On the Move to Meaningful Internet Systems: OTM 2019 Conferences; Confederated International Conferences; 2019 Oct; Rhodes, Greece. Springer, Cham. pp 714-30.

[3] Bonatti PA, Kirrane S, Petrova, IM et al. Machine Understandable Policies and GDPR Compliance Checking. Künstl Intell. 2020 Jul;34:303-15.

[4] Ryan P, Crane M, Brennan R. Design Challenges for GDPR RegTech. In: Filipe J, Smialek M, Brodsky A, Hammoudi S, editors. Proceedings of the 22nd International Conference on Enterprise Information Systems. ICEIS 2020; 2020 May; Prague, Czech Republic. SCITEPRESS 2020; 2020. p. 787-95.

[5] Debruyne C, Pandit HJ, Lewis D. et al. "Just-in-time" generation of datasets by considering structured representations of given consent for GDPR compliance. Knowledge and Information Systems. 2020 Apr;62:3615–40.

[6] Navigli, R, Ponzetto, SP. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence. 2012 Dec;193:217-50.

[7] Harkous H, Fawaz K, Lebret R, et al. Polisis:Automated analysis and presentation of privacy policies using deep learning. In: 27th USENIX Security Symposium. USENIX Security '18; 2018 Aug; Baltimore, USA. USENIX Association; 2018. p. 531–48.

[8] Wilson S, Schaub F, Dara AA, et al. The creation and analysis of a website privacy policy corpus. In: Erk K, Smith NA, editors. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2016 Aug; Berlin, Germany. Association for Computational Linguistics; 2016. p. 1330-40.

[9] Nejad NM, Jabat P, Nedelchev R et al. Establishing a strong baseline for privacy policy classification. In: Hölbl M, Rannenberg K, Welzer T, editors. ICT Systems Security and Privacy Protection. SEC 2020; 2020 Sep; Maribor, Slovenia. Springer, Cham. p. 370-83.

[10] Tesfay WB, Hofmann P, Nakamura T et al. PrivacyGuide:Towards an implementation of the EU GDPR on internet privacy policy evaluation. In: Verma RM, Kantarcioglu M, editors. Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics. Eighth ACM Conference on Data and Application Security and Privacy; 2018 Mar; Tempe, USA. Association for Computing Machinery; 2018. p. 15–21

[11] Sarne D, Schler J, Singer A et al. Unsupervised topic extraction from privacy policies. In: Liu L, White R, editors. WWW '19: Companion Proceedings of The 2019 World Wide Web Conference. The Web Conference; 2019 May; San Francisco, USA. Association for Computing Machinery; 2019. p. 563–68.

[12] Nejad NM, Scerri S, Lehmann J. Knight: Mapping privacy policies to GDPR. In: Faron Zucker C, Ghidini C, Napoli A, et al. editors. Knowledge Engineering and Knowledge Management. EKAW; 2018 Nov; Nancy, France. Springer, Cham; 2018. p. 258–72.

[13] Montiel-Ponsoda E, Gracia J, Rodriguez-Doncel V. Building the legal knowledge graph for smart compliance services in multilingual Europe. In: Rodríguez-Doncel V, Casanovas P, González-Conejero J, editors. Proceedings of the 1st Workshop on Technologies for Regulatory Compliance; 30th International Conference on Legal Knowledge and Information Systems; 2017 Dec; Luxembourg. CEUR Wordshop Proceedings; 2017. p. 15–17.

[14] Palmirani M, Bincoletto G, Leone V, et al. Hybrid Refining Approach of PrOnto Ontology. In: Kő A, Francesconi E, Kotsis G, et al., editors. Electronic Government and the Information Systems Perspective. EGOVIS; 2020 Sep; Bratislava, Slovakia. Springer, Cham; 2020. p. 3-17.

[15] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information processing & management. 1988 Aug;24(5):513-23.

[16] Kenter T, de Rijke M. Short Text Similarity with Word Embeddings. In: CIKM '15: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. CIKM; 2015 Oct; Melbourne, Australia. Association for Computing Machinery; 2015. p. 1411–20.

[17] Pennington J, Richard S, Manning C. Glove: Global vectors for word representation. In: Moschitti A, Pang B, editors. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP; 2014 Oct; Doha, Qatar. Association for Computational Linguistics; 2014. p. 1532-43.