Improving Readability Of Online Privacy Policies Through DOOP: A Domain Ontology For Online Privacy

Dhiren A. Audich [a,*], Rozita Dara [a] and Blair Nonnecke [a]
[a] *School of Computer Science, University of Guelph, Ontario, Canada*
E-mails: daudich@uoguelph.ca, drozita@uoguelph.ca, nonnecke@uoguelph.ca

Abstract. Privacy policies play an important part in informing users about their privacy concerns by operating as memorandums of understanding (MOUs) between them and online services providers. Research suggests that these policies are infrequently read because they are often lengthy, written in jargon, and incomplete, making them difficult for most users to understand. Users are more likely to read short excerpts of privacy policies if they pertain directly to their concern. In this paper, a novel approach is proposed that reduces the amount of text a user has to read. It does so by using a domain ontology and natural language processing (NLP) to identify key areas of the policies that users should read to address their concerns and take appropriate action. By using the ontology to locate key parts of privacy policies, average reading times were substantially reduced from $8-12$ minutes to 45 seconds.

Keywords: online privacy, privacy policy, ontology, natural language processing, human-computer interaction

1. Introduction

Many of the online activities place the cost of privacy on the users by requiring them to disclose their personal data in exchange for services. Due to the near permanent nature of the Internet, it results in a loss of privacy and can have a long-lasting effect on the user. A 2015 Pew Research Centre survey found that 91% of American adults either agree or strongly agree that they have lost control of how their private information is collected and used [1]. The collection of personal data by online service providers is often justified with claims of creating a more user-centric web experience. However, personal data is sold and shared frequently with third parties that use it to profile users and track them across domains. Many surveys and studies have suggested that users are increasingly concerned about their privacy online [2]. To ease user concerns and bolster trust, companies are introducing privacy enhancing technologies (PET) such as: opt-out mechanisms; reducing the amount of personal information collected; anonymization of personal data; and 'layered' policies [3, 4]. Without these becoming a common standard, opaque and verbose policies are still the norm.

Privacy polices offer a glimpse into how user data is collected and disseminated. They are designed to reduce fear among users concerning their personal information [5]. By law, privacy policies are required to disclose the nature and extent of information collection [6–9]. Unfortunately, most policies are often lengthy, difficult and time-consuming to read, and as a result are infrequently read [2, 10–12]. The demotivating nature and the difficulty of reading privacy policies amounts to a lack of transparency. Failing to provide usable privacy policies prevents users from making informed decisions and can lead them to accept terms of use jeopardizing their privacy and personal data.

This paper proposes a novel approach to reduce the amount of text a user has to read by using a domain ontology to identify key areas that address their concerns and allow them to take appropriate action. The approach consists of constructing a domain ontology for

[*] Corresponding author. E-mail: daudich@uoguelph.ca.

online privacy (DOOP), an ontology for online privacy policies, validated against Carnegie Mellon University's (CMU) OPP-115 data set [13] of annotated policies by domain experts. DOOP resulted in $69\% - 99\%$ reductions in reading for the three sample queries that were tested. Reducing the reading time will encourage users to read privacy policies and make informed decisions online.

It is important to note that DOOP is the first ontology to capture the vocabulary of online privacy policies. It also provides a method to describe the vocabulary in terms of the privacy categories that are widely used by Federal Trade Commission and directives proposed by other commissions in Europe and Canada.

2. Background

To aid users, several attempts have been made to simplify policies. One of the early efforts was the Platform for Privacy Preferences (P3P) [14, 15], which introduced a machine readable format for creating privacy policies. The intent was that the standardized format would make it easier to extract relevant information with the help of logic systems, e.g., reasoners. P3P had limited success due to a lack of industry and developer participation. It also lacked proper policy validation which prevented policy developers from creating accurate policies [16].

Automation and crowdsourcing can reduce the cost of creating and maintaining such a data set, and still maintain reasonable quality. Terms of Service; Didn't Read (ToS;DR) [17] is a project that uses crowdsourced annotations to answer key questions about the policies. The limiting factor with crowdsourcing is the large scale participation rate required to ensure success, leading to delays in the success of the project. To remedy this, researchers [18, 19] tried to combine ToS;DR, natural language processing (NLP) and other machine learning techniques with the goal of automatically inferring privacy concerns from privacy policies. Whilst these techniques work well in recognizing the pre-determined classes of privacy concerns, they still rely on quality, reliable, and up-to-date crowdsourced data which is presently lacking due to the inherent demotivating nature of reading privacy policies [18].

In the research conducted by Ramnath et al., the researchers proposed combining machine learning and crowdsourcing (for validation) to semi-automate the extraction of key privacy practices [20]. Through their preliminary study they were able to show that non-domain experts were able to find an answer to their privacy concern relatively quickly ($\sim$ 45s per question) when they were only shown relevant paragraphs that were mostly likely to contain an answer to the question. They also found that answers to privacy concerns were usually concentrated rather than scattered all over the policy. This is an important find because it means that if users are directed to relevant sections in the policy they should be able to address their privacy concerns relatively quickly.

In a more recent user study conducted by Wilson et al. (2016), the quality of crowdsourced answering of privacy concerns was tested against domain experts with particular emphasis on highlighted text. The researchers found that highlighting relevant text had no negative impact on accuracy of answers. They also found out that users tend not to be biased by the highlights and are still likely to read the surrounding text to gain context and answer privacy concerning questions. They also found an 80% agreement rate between the crowdsourced workers and the domain experts for the same questions [21]. These findings suggest that highlighting relevant text with appropriate keywords can provide some feedback to users inclined to read shorter policies. One way to automatically highlight relevant text in privacy policies, in a manner that can be easily scaled, is by using semantic technologies (ST) such as an ontology.

3. Motivation

As established in the introduction, online privacy policies remain unusable to the average users due to their length and elusive language. This contributes to a lack of transparency which in turn leads to uninformed decisions and risk-averse behaviour. Policies that are usable tend to be read more often and give the users more confidence in sharing their personal information. This suggests that a usable privacy policy benefits all parties. Research shows that policies which highlight sections that directly address the user's concerns tend to be read more often, as it reduces the reading cost. Hence, a solution is required which considers concerns of all stakeholders, i.e., the online service providers that create privacy policies, as well as the users that do not necessarily like reading them. To avoid pushback, the solution must not require the online service providers to change their policies drastically, but must reduce the amount of text users have to read, and direct users to the text pertaining directly to their concerns.

In order to direct users to the relevant text within the policies, there needs to be a way to evaluate the text and highlight all relevant sections. Since the language being used within the policies differs so greatly [11, 19], there needed to be a system that is able to capture all the variations of a topic. Semantic technologies such as ontologies are a well-known way of mining text and reasoning. Since domain ontologies can capture the vocabulary of a domain and specify rules about each term, it is possible to capture the diversity of concepts within online privacy domain and reason over them to find equivalent terms for analysis. Through NLP, it is possible to logically break apart the text in a policy, and working in conjunction with the ontology, be able to recognize relevant sections within policies. This paper proposes building an ontology that is easy to build, maintain, and relatively inexpensive to scale.

4. Ontology Engineering

It is generally accepted that ontologies have two basic features [22]:

1. A taxonomy of terms used to name and describe the objects (concepts) being described by the ontology.
2. A specification, grounded in logic, used to add meaning between terms.

Ontologies may describe a wide variety of things in a domain, but they all share a common set of attributes [23]:

- **Classes** capture the core vocabulary that is used to describe a domain. They are also referred to as *concepts*, and are generally arranged in a hierarchical or taxonomical form as classes and subclasses.
- **Relations** are definitions of how concepts interrelate to one another.
- **Attributes** are the properties associated with classes that describe the features of that class.
- **Formal axioms** are logical statements that always evaluate to true.
- **Functions** are a special case of relations.
- **Instances** are elements of a class; and are also called *individuals*. Not all ontologies must have these; but if they do then that ontology constitutes a *knowledge base*.

There are many different types of ontologies that differ based on not only their purpose but also their content. The purpose of the ontology is determined by how widely it is meant to be used and the content is determined by the richness of the term definitions.

Since the invention of ontologies in the early 90s, ontology engineering has remained more of an art form rather than an engineering process with rigid rules [24]. Which is to say, there is no one correct way of creating an ontology, rather the development differs depending on the ontology engineer and its purpose. However, several methodologies have been proposed to standardize the process of creating ontologies. Among those, Ontology 101 [25] and NeOn [26] are two of the commonly used ontology engineering methods. Ontology 101 provides a step-by-step methodology for creating simple ontologies iteratively using the Protégé-2000 [27] ontology engineering tool, developed by Mark Musen's research group at Stanford Medical Informatics. Since this methodology is geared towards the Protégé tool, it focuses more on declarative frame-based systems used to describe objects in a domain along with their properties rather than more complex and domain specific ontologies that can be constructed with the other methodologies. The NeOn methodology promotes reusing and combining ontologies to create new and networked ontologies drawing on multiple ontologies for their knowledge. NeOn provides a scenario-based framework to create ontologies and develop and expand ontology networks. Rather than a rigid work flow like Ontology 101, NeOn prescribes a set of guidelines for multiple alternative processes for various stages of ontology development that may change with the design decisions.

There are a few ways of evaluating an ontology, each depending on the type and purpose of the ontology constructed. One of the practical ways of evaluating ontologies is a data driven approach. In this approach, ontologies are simply compared with the sources of data from the domain that the ontology is meant to cover. This involves statistically extracting key terms and concepts from the corpus the ontology is meant to cover and evaluating if they exist in the ontology itself. This is done via calculation of the *precision* and *recall* scores [28–31].

Precision (P) is the ratio of the number of relevant terms returned from a term extraction algorithm ($\{\text{manually-selected}\}\cap\{\text{machine-selected}\}$) to the total numbers of retrieved terms by the algorithm ({machine-selected}). The precision is calculated using Equation 1.

$$P = \frac{\mid \{\text{manually selected}\} \cap \{\text{machine-selected}\} \mid}{\mid \{\text{machine-selected}\} \mid} \tag{1}$$

The *recall (R)* of an information system is defined as the ratio of the number of relevant terms returned to the total number of relevant terms in the collection. The recall is computed using Equation 2.

$$R = \frac{\mid \{\text{manually selected}\} \cap \{\text{machine-selected}\} \mid}{\mid \{\text{manually selected}\} \mid} \tag{2}$$

Formal competency questions (CQs), an evaluation strategy proposed by [32, 33], is another ontology validation method. In this method, informal competency questions (queries) are first expressed formally. These questions are requirements that are in the form of questions that an ontology must be able to answer. The formal questions are then evaluated using completeness theorems with respect to first (axioms) and second order (situational calculus) logic representation of concepts, attributes, and relations.

5. Methodology

The end goal of DOOP is the creation of a tool, e.g. a browser extension, that uses the ontology as a knowledge base to parse online privacy policies and highlight sections in the policy that would address a user's concerns. To that end, a hybrid construction approach was used, a combination of Ontology 101 and NeOn methodologies in conjunction with iterative development. Ontology 101 was proposed with the intent of using Protégé for the construction of ontology. Since the latest version of Protégé [34] was used in the construction of DOOP, Ontology 101 was used as the prime methodology. Furthermore, ontology engineering guidelines provided under NeOn's Scenario 1 (From Specification to Implementation), which includes steps from other methodologies, were used as the foundation to specify and build DOOP. Instead of building a complete ontology that exhaustively considers every possible case, DOOP was iteratively built in iterations as described by RapidOWL. By expanding the vocabulary one query at a time the ontology remains open and malleable enough such that future developments require relatively less effort to alter the structure of the ontology as needed. The ontology was implemented in OWL-DL using the Protégé tool (version 5.2.0) for ontology engineering.

The language between privacy polices is inconsistent. These inconsistencies meant that the corpus of privacy policies used for ontology development had to be large and diverse to capture as many terms as possible, and from different economic zones. For this reason, we took a two-step approach to extract keywords and validate the ontology. First, the key terms extracted from a corpus 631 privacy policies were gathered [35]. Seven classes were identified: data collection, data retention, data security, data sharing, target audience, user access, and user choice. These were identified in the work done by [36]; and were based on the logical division of policies described under the FIPPs (Fair Information Practice Principles), and principles identified under OECD's guidelines for (Organization for Economic Co-operation and Development) protection of privacy and data flows [37]. These classes are commonly found in both cookie and privacy policies. The hierarchy of the ontology was developed as needed to satisfy the CQs. Subsequently, CMU's OPP-115 corpus, a corpus of 115 manually annotated privacy policies with $23,000$ data practices was used to validate DOOP. Lastly, the goal of building DOOP was to support the following end-users: privacy researchers, NLP experts who are interested in doing work in the online privacy domain, and software developers who would like to use an ontology to create tools for the online privacy domain. An overview of DOOP construction and validation is shown in Figure 1.

5.1. Competency Questions

Competency questions are a set of queries that the ontology should be able to answer based on its axioms. This is why they are used for not only defining ontology requirements but also ontology evaluation; the result from a CQ can be used to determine the correctness of an ontology. CQs can be used for evaluation either manually or automatically through the use of SPARQL queries. DOOP was constructed and evaluated through the use of CQs. After defining a CQ, the ontology was constructed iteratively by defining as many axioms needed to answer the CQ. The following 3 CQs were used for constructing the ontology:

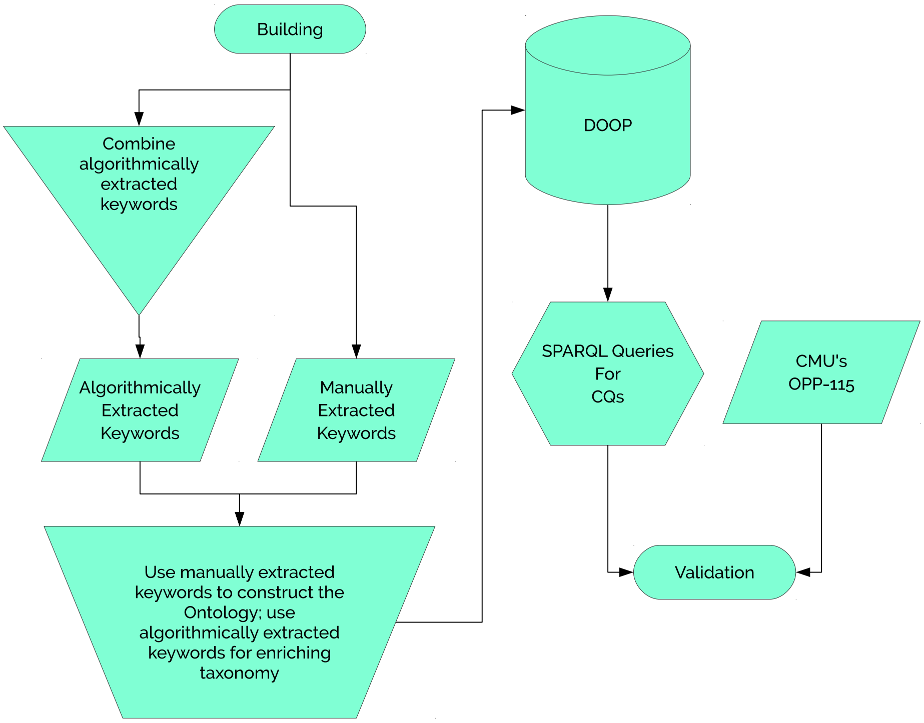

Building
Combine algorithmically extracted keywords
Algorithmically Extracted Keywords
Manually Extracted Keywords
Use manually extracted keywords to construct the Ontology; use algorithmically extracted keywords for enriching taxonomy
DOOP
SPARQL Queries For CQs
CMU's OPP-115
Validation

Fig. 1. Overview of how DOOP was constructed and validated.

1. Does this website share my personal information with third-parties?
2. Does this website use tracking cookies?
3. Can I opt-in/opt-out of information gathering?

The first query was selected based on the most common concern users report, a worry/concern about their personal information being shared with unauthorized and unintended parties. It is also a question that every privacy policy is designed to address. The other two queries are more specific and hence were chosen to capture narrower results.

6. Results

A breakdown of DOOP is given in Table 1. An example of a class is given in Table 2.

6.1. Top structure

The ontology is divided into four main classes derived from the standard OWL root class for everything *owl:Thing*: Keyword, PrivacyCategory, PrivacyPolicy, and Question. The following are rules for creating the rest of the classes and individuals. Refer to Figures 2 and 3.

Table 1

Breakdown of DOOP.

Property	**Count**
Axioms	304
Logical axiom count	171
Declaration axioms count	71
Class count	15
Object property count	6
Data property count	10
Individual count	35
Annotation property count	8

1. The Keyword class captures most of the vocabulary contained in DOOP. As shown in Figure 2, it is a sub-class of *owl:Thing*. Classes act as sets of individuals, hence, Opt-in and Opt-out form instances of Keyword (Fig. 3) but Cookie is a sub-class of Keyword; this is because the class Cookie has the instances, ‘Do Not Track’ and ‘Web Beacon’, which are **types** of cookies. Legal Act, Country, and Organization, are all sub-classes of Keyword which are interrelated by relationships: ‘Applies To’, ‘Operates In’, ‘Enacted’, and ‘Based’. Legal acts ‘Applies To’ country which is an inverse of ‘Enacted’. Similarly, or-

Table 2

Example of a class in DOOP.

Preferred Name	Personally Identifiable Information
Alternative	PII
Description	Information about an individual person, organization, or any other entity.
SubClass Of	Keyword
Instances	Email Address, Email Preference, Name, Age, PII

ganizations 'Operates In' countries, and countries serve as 'Base' for organizations which is also an inverse relationship to the former relationship.

2. Privacy category class does not have a sub-class, but has 7 individuals: data collection, data retention, data security, data sharing, target audience, user access, and user access. Both Keyword and Question classes share a 'Related To' relationship with Privacy Category, since keywords and questions are logically classified under various privacy categories. An example of this is shown in Figure 3, where Opt-Out is 'Related To' User Choice, Data Collection, and Data Sharing. Since, Opt-Out is 'Similar To' Opt-In, it is inferred that it too is related to the three categories.
3. The Question class is where all of the questions are stored as individuals. This class has no subclass because all of the questions are **types** of Question. Additional information is stored as object properties, e.g., concern, competency question, and SPARQL query.
4. The Privacy Policy class stores meta-data for privacy policies already processed. It is divided into two sub-classes: Policy Document and Cookie Policy. Privacy policies are individuals of the Policy Document class, and if a seperate coolie policy exists then its meta-data is stored under the Cookie Policy class.

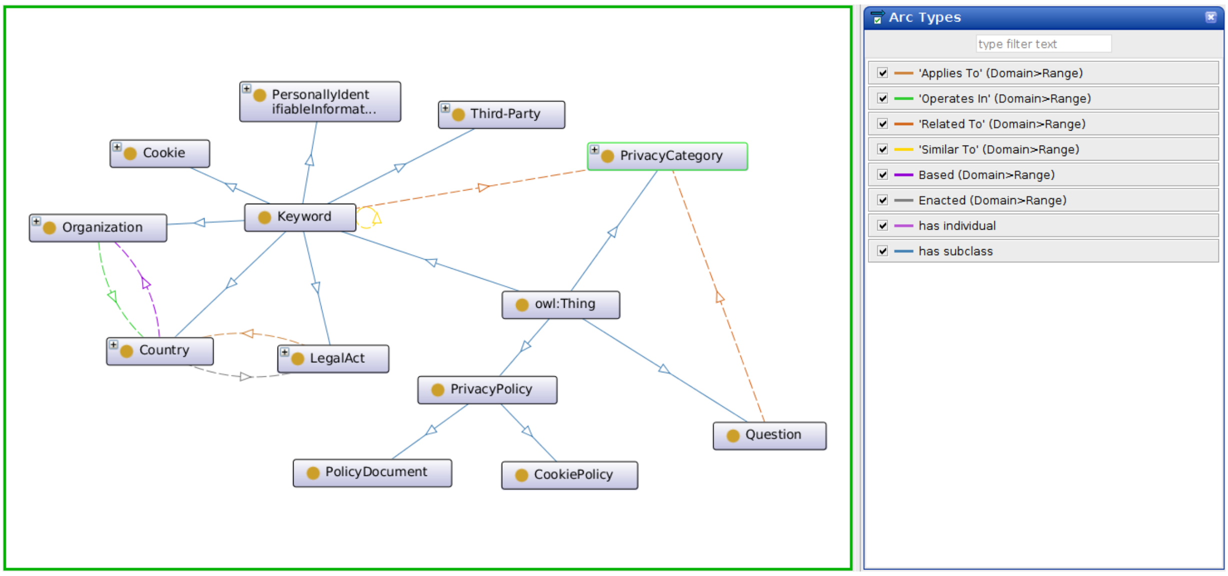
The general structure of the ontology was developed with the firm intention of developing a tool to dissect an online privacy policy into sections that the user might be most interested in based on their concerns and the captured vocabulary offered by the domain ontology. Since, to the best of our knowledge, no previous attempt had been made to capture only the vocabulary in online privacy policies, there were no ontologies to refer to for creating classes and structure.

6.2. *Inference and structure*

DOOP is primarily composed of single *is-a* asserted inheritance structure, expressed with subclass relations in OWL-DL. However, other relations also exist to enable further development and capturing of more complex logic. An exhaustive list of object relations along with their properties is described in Table 3. These provide useful classification hierarchies and extendability for the users of the ontology. These relationships allow the user to infer which keywords are related to what privacy category and question; which questions share a certain number of keywords, useful for recommendation; what legal acts are enforceable in a country; what organizations enforce which legal acts and where they are located; determine the overlap of vocabulary between privacy categories. DOOP is consistent with all three reasoners in Protégé 5.2.0: FaCT++ [38], HermiT [39], and Pellet [40].

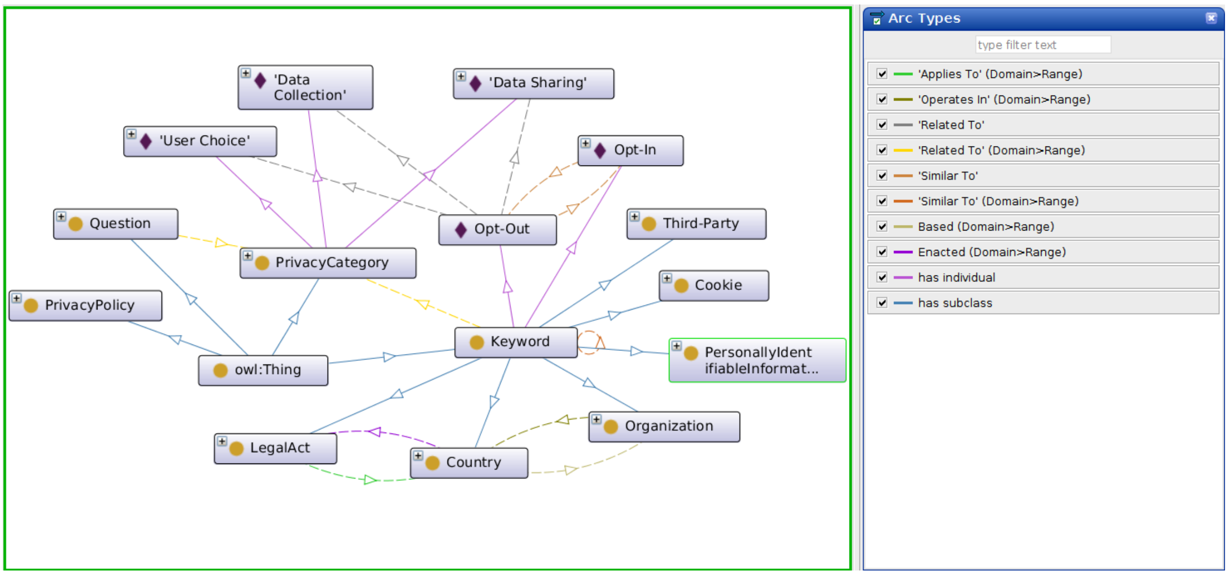
6.3. *Validation*

Owing to the individual and subjective nature of ontologies, ontology validation is a difficult process. DOOP was validated in two ways: CQs and data driven. CMU's OPP-115 data set was used as the primary benchmark for validation. Since there were 10 policy annotators, there were many overlaps in the labelling of policies, as well as multiple labels for the same policy. To reduce redundancy, the authors of OPP-115 consolidated annotations with three convergence thresholds: 0.5, 0.75, and 1. These were calculated as normalized aggregated overlap of spans of text assigned the same data attribute for the same data practise identified by multiple annotators. For our validation, annotations from all three thresholds were considered. Furthermore, OPP-115 annotators have classified their annotation under 10 categories which had to be mapped to our 7 categories; Figure 4 shows this mapping. The mapping was necessary because at the time of undertaking this research OPP-115 data set had not yet been released so the categories introduced by [36] were used. For validation, four experiments were conducted that evaluated various aspects of the ontology:



☑ 'Applies To' (Domain>Range)
☑ 'Operates In' (Domain>Range)
☑ 'Related To' (Domain>Range)
☑ 'Similar To' (Domain>Range)
☑ Based (Domain>Range)
☑ Enacted (Domain>Range)
☑ has individual
☑ has subclass

Fig. 2. Top class structure of DOOP. Some classes and individuals are not shown due to space limitations.



☑ 'Applies To' (Domain>Range)
☑ 'Operates In' (Domain>Range)
☑ 'Related To'
☑ 'Related To' (Domain>Range)
☑ 'Similar To'
☑ 'Similar To' (Domain>Range)
☑ Based (Domain>Range)
☑ Enacted (Domain>Range)
☑ has individual
☑ has subclass

Fig. 3. Example of individuals in DOOP. Some classes and individuals are not shown due to space limitations.

1. **Correctness:** Compute the number of matched privacy categories for the same sentences from both DOOP (based on the keywords the sentence contains) and OPP-115.
2. **Policy coverage:** Compute the the number of sentences that the reader has to theoretically read to understand the risks associated with his concerns.
3. **Completeness:** Compute per policy keywords that existed in the ontology but not in the policy.
4. **Correctness:** Compute cases where the keyword's assigned category in the ontology did not match the OPP-115's annotation's assigned category.

6.3.1. Experiment 1: Correctness

In order to compare categories, the annotation set's results had to be processed. The results are presented as a CSV file for each privacy policy in the `consolidation` directory as indicated by their

Table 3

Object properties presently available in DOOP.

Relationship	Domain	Range	Characteristics
is a	∞	∞	Transitive
superclass of	∞	∞	
applies to	LegalAct	Country	Functional, InverseOf:enacted
based	Country	Organization	InverseOf:OperatesIn
enacted	Country	LegalAct	InverseOf:AppliesTo
operates in	Organization	Country	Functional, InverseOf:Based
related to	Keyword, Question	PrivacyCategory	Symmetric
similar to	Keyword	Keyword	Transitive, Symmetric

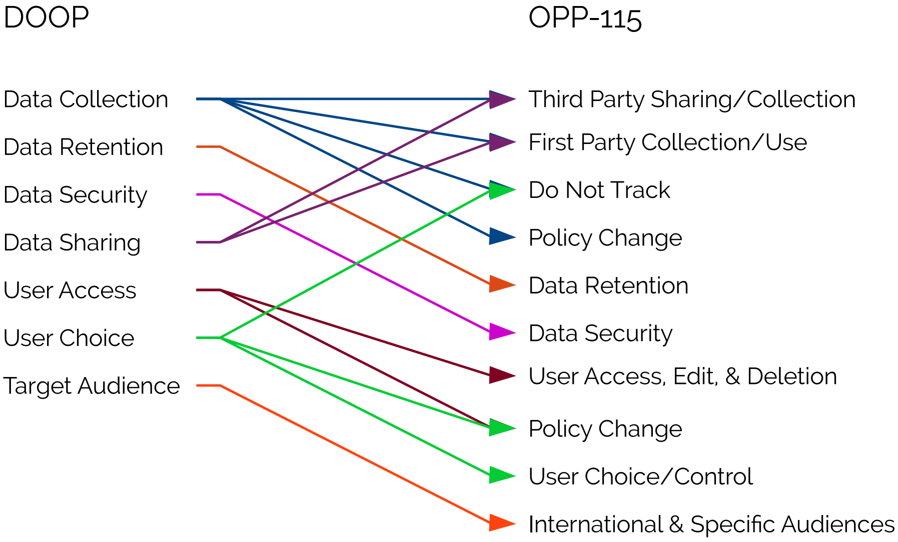


Fig. 4. Mapping DOOP categories to OPP-115.

manual. The column description used by the CSV files is presented below:

(A) annotation ID (a globally unique identifier for a data practice)
(B) batch ID (name of a batch in the annotation tool; often indicates who the annotators were)
(C) annotator ID
(D) policy ID (this corresponds to the numeric prefixes in the policy filename, as found in other directories)
(E) segment ID (the zero-indexed, sequential identifier of the policy segment; e.g., the first segment in a policy's text is segment zero)
(F) category name
(G) attribute-value pairs (represented as JSON, this where the annotations are stored)
(H) policy URL
(I) date

A qualified positive match of categories occurs when the `selectedText` for an annotation under `attribute-value` contains any of the keywords returned by DOOP for a query whose DOOP categories also match their mapped OPP-115 `category name`. Results for all three queries, and for all three thresholds are presented in Table 4. A sentence in a policy that contains any of the keywords that DOOP returns is denoted by SM or Sentences Matched, and sentences for which there is category match is denoted by PM or Positive Matched, while score is the percentage.

6.3.2. *Experiment 2: Coverage*

For this experiment, the average number of sentences that exist in a policy was first calculated, then it was divided by the number of sentences that contained

Table 4
Results for DOOP validation, experiment 1.

Query	0.5			0.75			1		
	SM	PM	%	SM	PM	%	SM	PM	%
1	5254	3262	**62.09%**	5926	3689	**62.25%**	6540	3931	**60.11%**
2	322	289	**89.75%**	368	331	**89.95%**	401	362	**90.27%**
3	803	622	**77.46%**	907	713	**78.61%**	1012	759	**75.00%**

the keywords that were returned after the execution of the SPARQL query for a privacy concern. The results for all queries as well as for all convergence thresholds are presented in Table 5. Average number of total sentences in privacy policies is denoted by TS, whereas selected sentences that contain a keyword also in the returned query from DOOP is denoted by SS.

6.3.3. *Experiment 3: Completeness*

In this experiment, the number of keywords that the ontology returned for a particular query that did not exist in the privacy policies was calculated. The primary purpose here was to investigate how many unique terms existed in DOOP that did not exist in the policy. Since all of 115 policies used in OPP-115 were American, unique terms found in DOOP would indicate a geographic non-specific ontology that can be generally used in most English speaking countries. Results from this experiment are reported in Table 6. Average number of keywords found is denoted by KF, and average number of keywords not found is denoted by NF.

6.3.4. *Experiment 4: Correctness*

Since, assigning categories to keywords is a manual task, and privacy categories from DOOP were further mapped onto category names assigned by annotators from OPP-115, we wanted to know how often we differed in opinion. Thus, in this experiment we investigated how often keyword assigned category in DOOP differs from OPP-115. Results from this experiment is presented in Table 7.

7. Discussion

In Experiment 1 (Table 4), a mean of 76.16% match for privacy categories with a standard deviation of 12.41 was achieved. Since the OPP-115 data set was manually curated by domain experts, a high degree of match indicates a high degree of accuracy achieved by the ontology for identifying sentences in context based on the vocabulary. Now, the users need not read the entire policy, but can be directed to appropriate sentences in the policy that deal directly with their concerns with a reasonable amount of accuracy. Additionally, there is a negligible increase in the accuracy of the automatic categorization when the convergence threshold is 0.75. This could be as consolidation reduces redundancy, without overdoing it at 1.0 convergence threshold.

One of the prime reasons that users do not read privacy policies and are left uninformed, is that they tend to be overly long. Any tool trying to fix this issue must not only find correct information but also require less reading. In Experiment 1, algorithmic assignment of privacy categories to sentences performed favourably against the manual annotations performed by domain experts. Thus, in Experiment 2, policy coverage was investigated to identify how much of a policy is the user asked to read for the three identified concerns. This experiment demonstrated (Table 5) that the user has to read on average 11.09 sentences with a standard deviation of 14.70, or about 12.09% of a policy with a standard deviation of 16.06 to know if all of their concerns are met. Assuming that a paragraph is roughly 10 sentences, then based on research done by [20], we know that it would take roughly 45 seconds to read it. This reduced time makes the privacy policies more inviting and should encourage more users to read policies even if partially.

To further investigate these results, additional analysis of the individual results from experiments for all of the thresholds and queries was conducted. Figures 5, 6, and 7 show the results of the queries for the 0.75 threshold. A radical relationship between the number of sentences selected and the length of the policies was expected, where the reading proportionally increases as the length of the policies increase, but then stabilize at some horizontal asymptote. However, this did not occur. A strong positive linear relationship was observed for the first experiment (Fig. 5), no correlation for the second experiment (Fig. 6), and weak positive

Table 5

Results for DOOP validation, experiment 2.

Query	0.5			0.75			1		
	TS	SS	Coverage	TS	SS	Coverage	TS	SS	Coverage
1	91.7	27.99	**30.99%**	91.7	27.99	**30.99%**	91.7	27.99	**30.99%**
2	91.7	1.18	**1.42%**	91.7	1.18	**1.42%**	91.7	1.18	**1.42%**
3	91.7	4.12	**4.32%**	91.7	4.12	**4.32%**	91.7	4.12	**4.32%**

Table 6

Results for DOOP validation, experiment 3.

Query	0.5			0.75			1		
	KF	NF	Unique	KF	NF	Unique	KF	NF	Unique
1	3.86	9.14	**70.31%**	3.86	9.14	**70.31%**	3.86	9.14	**70.31%**
2	0.41	2.59	**86.33%**	0.41	2.59	**86.33%**	0.41	2.59	**86.33%**
3	1.39	1.61	**53.67%**	1.39	1.61	**53.67%**	1.39	1.61	**53.67%**

Table 7

Results for DOOP validation, experiment 4.

Query	0.5			0.75			1		
	Total	Diff	%	Total	Diff	%	Total	Diff	%
1	5254	1992	**37.91%**	5926	2237	**37.75%**	6540	2609	**39.89%**
2	360	78	**21.67%**	415	91	**21.93%**	465	110	**23.66%**
3	1606	984	**61.27%**	1814	1101	**60.69%**	2024	1265	**62.50%**

for the third experiment (Fig. 7). A qualitative analysis provided several clues for these behaviours:

1. The vocabulary in the first query, was trying to capture more than one concern. Since the ontology only returned a vocabulary of terms, it was hard to determine the correct context sometimes as one set of keywords could be used in multiple instances under different contexts. One possible solution to this problem is having the ontology also capture POS tags that determines the structure of the sentences and identifies the associated verb (e.g. sharing) and thus provide a context under which the sentence occurs. This would help distinguish one context from another where the most of the vocabulary is shared. This idea is explored in Section 7.2.
2. Policies were repeated throughout the document. This accounted for the linear relationship for the first experiment. Redundancies in the policies drove up the number of sentences to read.
3. Keywords returned by the ontology were narrowly defined. This was an important distinction with the second query regarding tracking cookies. The keyword ‘cookie’ was not being used because not all cookies are tracking, this meant that several cases where that term was being used to establish context were not captured. For example, “We do not use tracking cookies.” would be selected, but, ‘Cookies may be used to track you’, would be ignored. Similar to the first problem, POS tags for some terms could be captured by the ontology to identify context as a remedy to this problem. In the OPP-115 data set, only 30% of the policies had a ‘tracking cookie’ policy that was part of the privacy policy that was extracted.
4. Some policies were missing entirely. Sometimes, a supplementary document was used to state policies, e.g., ‘Cookie Policy’. This supplementary document was not stored on the same page as the privacy policy; hence, it was not picked up by the scraper scripts. This a difficult challenge to solve as there is no consensus as to what the URL must be for the cookie policy. However, a reasonable

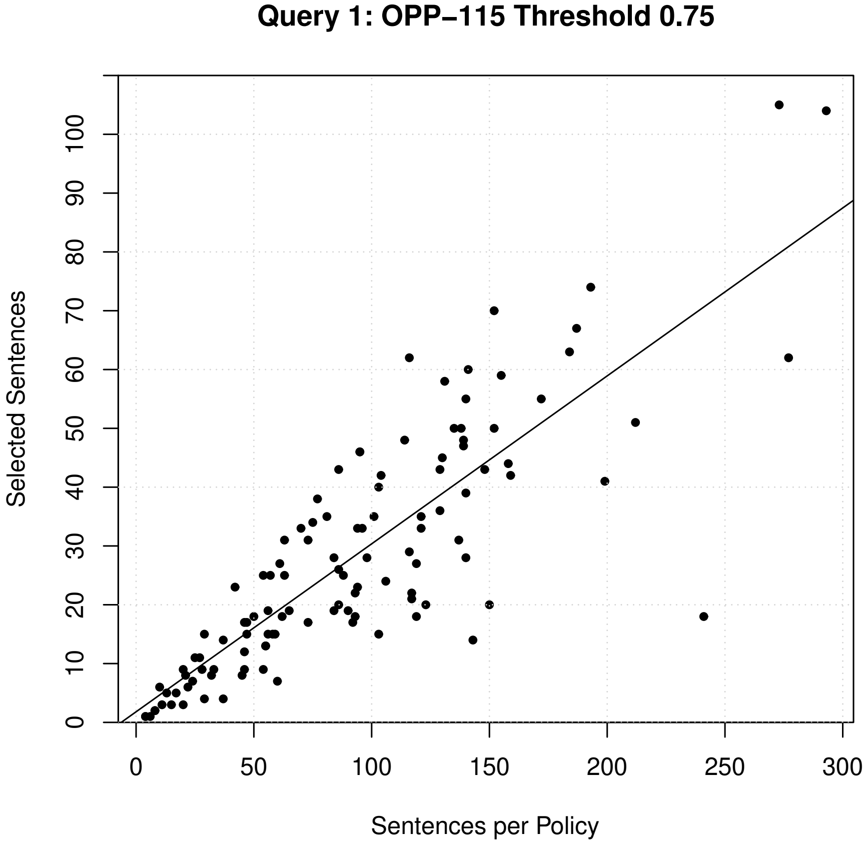


Fig. 5. Sentences selected for reading for query 1: 'Does this website share my personal information with third-parties?'.

attempt can be made to collect this page as well and amend it to the privacy policy.

In the creation of the taxonomy for DOOP, the vocabulary was not restricted to a particular geographically intended audience (in order to make the ontology as general purpose as possible). OPP-115's data set contained only American privacy policies. Hence, there were terms in DOOP that did not exist in OPP-115. Experiment 3 was conducted to investigate the uniqueness of DOOP in comparison to OPP-115. Table 6 shows that on average 70.10% of terms are unique to DOOP with a standard deviation of 14.14. This was expected as not just American policies were considered when extracting keywords from privacy policies, but also Canadian and European ones. The 70% terms also include localization of the American spelling along with synonyms, hypernyms, and E.U. and Canada specific terms, which made the ontology more unique here, e.g., advertiser/advertizer, and name/full name.

Finally, Experiment 4 investigated how much the labels of keywords agreed between DOOP and OPP-115. The mean disagreement between the data sets was 40.81% with a standard deviation of 17.03. The most disagreement being with query 3. One of the reasons for this discrepancy could be due to the mapping of categories from OPP-115 to DOOP. Since the mapping of the privacy categories between the data sets was not one-to-one, approximations had to be made. This meant that one category in one data set was mapped onto multiple categories in the other introducing a large amount of variance in the topics captured by each category. Another explanation has to do with the limited vocabulary DOOP currently captures. In its present state it was created to be a proof-of-concept system. As the vocabulary increases, it is expected that the results for all four experiments will also improve.

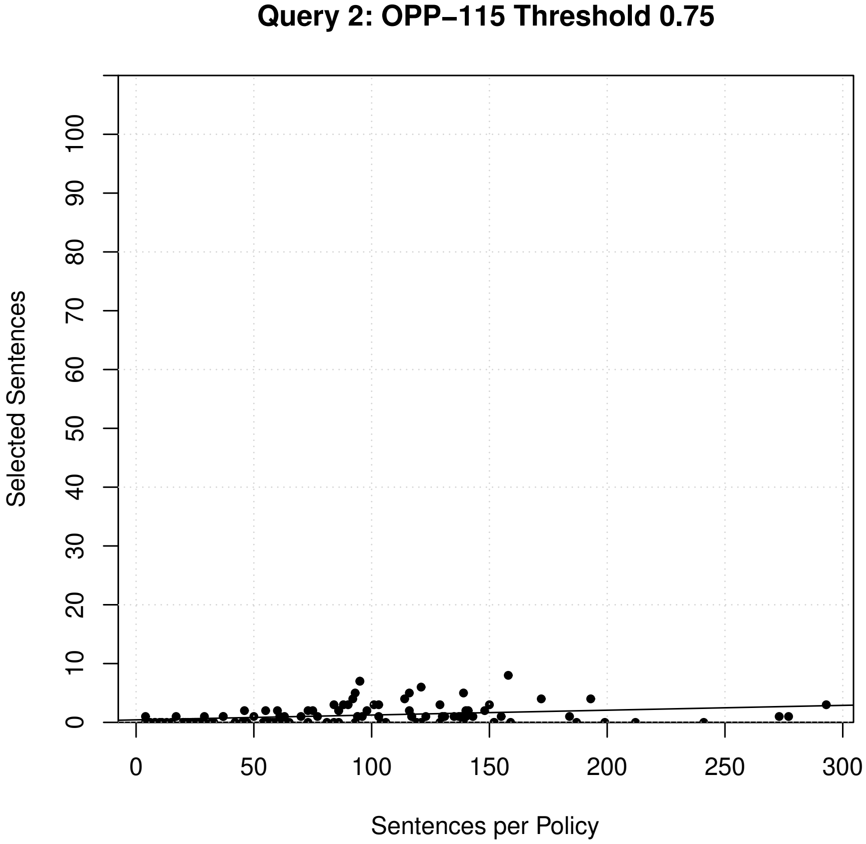


Fig. 6. Sentences selected for reading for query 2: 'Does this website use tracking cookies?'.

7.1. Generalizing keyword searches

Investigation was conducted to see what happens when more generic keywords are added to queries for focused and narrowly defined queries, such as query 2. The term 'cookie' was added to the list of keywords for the second query, and all of the experiments re-run. The results are shown in Tables 8, 9, 10, 11 and Figure 8.

In general, the total number of sentences dramatically increased from 364 to 1503 (Table 8), and the accuracy went from 90% to 85%. This was expected as 'cookies' was mentioned more often because they are used for more than just tracking. They are also widely used for storage of temporary data. This can be also observed in Table 10 which shows there is at least one word common to the vocabulary in the ontology and is consistently being found in the policies. Furthermore, this resulted in an increase of the estimated number of sentences to read per policy (on average going from 1.18 to 8.12 Table 9). The addition of non-specific keyword also increased the variability of the sentences to read, as can be observed in Figure 8. This also led to fewer policies being flagged as having 0 sentences to read. Once again, the number of sentences in a policy and the amount of reading a user has to do is linearly correlated. One of the most important measure is the increased disagreement between the recommended and the annotated sentences (Table 11). This indicates that adding generic terms deteriorates the overall quality of the recommendations. The indicators in this short study demonstrate that in order for the recommended reading to be useful to the user, it must have fewer generic terms and more targeted ones.

7.2. Contextualising keyword searches

Upon qualitative review of the sentences that were selected for Query 1 in Experiment 2, it was found

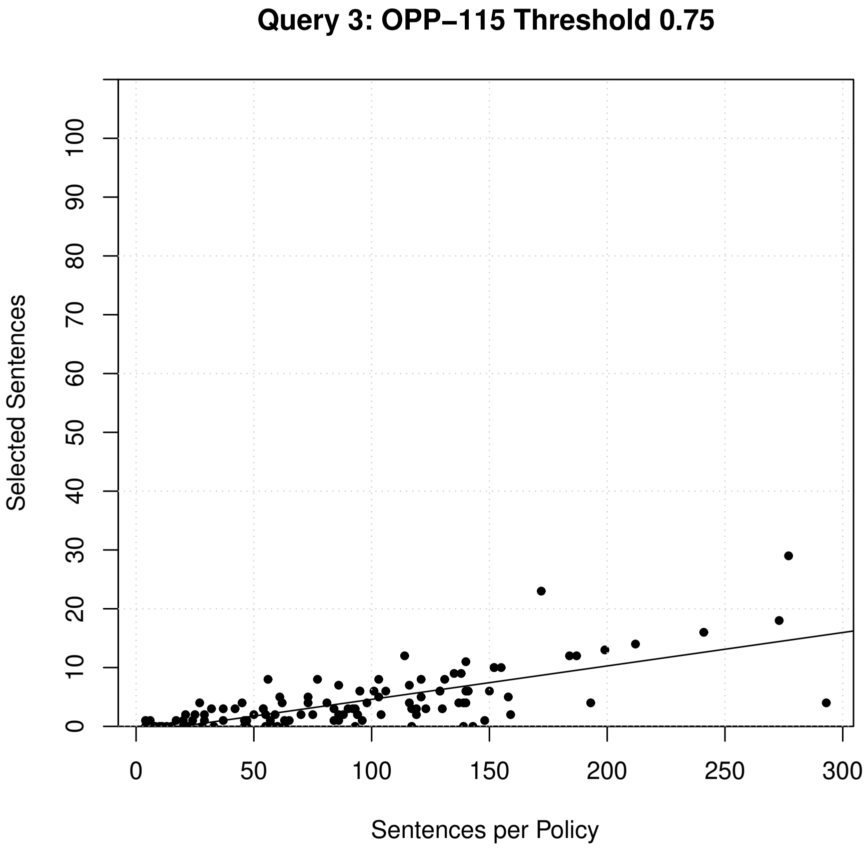


Fig. 7. Sentences selected for reading for query 3: ‘Can I opt-in/opt-out of information gathering?’.

Table 8
Results for DOOP validation for query 2, experiment 1.

‘Cookie’	0.5			0.75			1		
	SM	PM	%	SM	PM	%	SM	PM	%
No	322	289	**89.75%**	368	331	**89.95%**	401	362	**90.27%**
Yes	1367	1157	**84.64%**	1516	1295	**85.42%**	1625	1380	**84.92%**

Table 9
Results for DOOP validation query 2, experiment 2.

‘Cookie’	0.5			0.75			1		
	TS	SS	Coverage	TS	SS	Coverage	TS	SS	Coverage
No	91.7	1.18	**1.42%**	91.7	1.18	**1.42%**	91.7	1.18	**1.42%**
Yes	91.7	8.12	**9.69%**	91.7	8.12	**9.69%**	91.7	8.12	**9.69%**

that sentences were broadly selected to capture any mention of attributes associated with personal information and not necessarily- third-party sharing. To reduce the number of sentences selected for reading

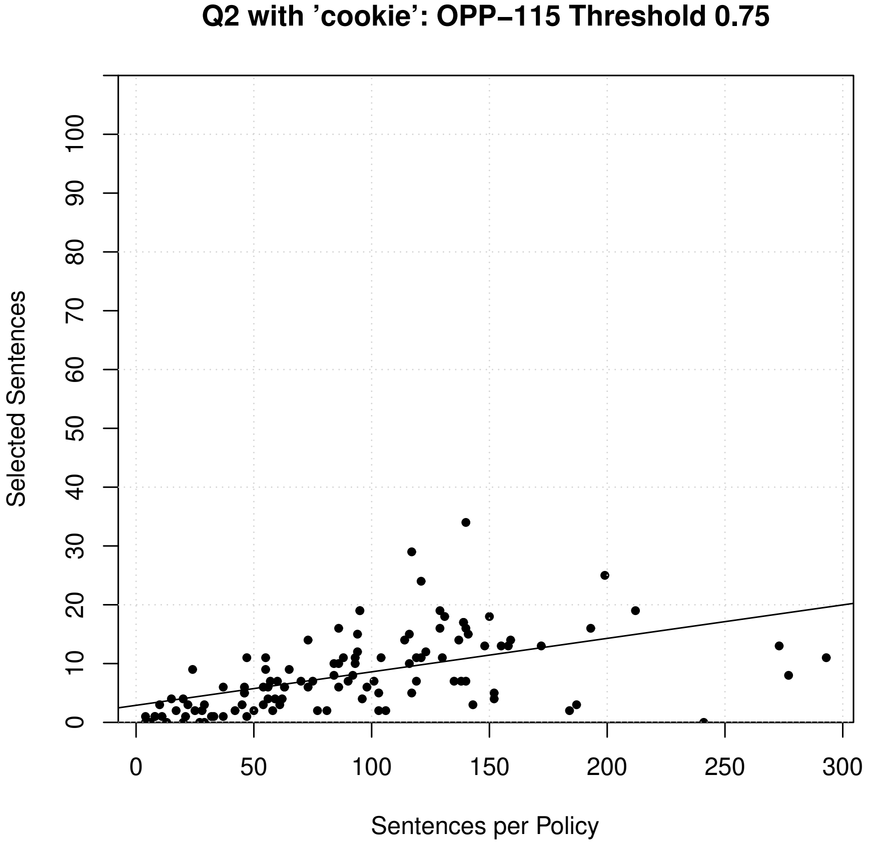


Fig. 8. Results for experiment 2 for query 2 after adding 'cookie' to the keywords.

Table 10
Results for DOOP validation for query 2, experiment 3.

'Cookie'	0.5			0.75			1		
	KF	NF	Unique	KF	NF	Unique	KF	NF	Unique
No	0.41	2.59	**86.33%**	0.41	2.59	**86.33%**	0.41	2.59	**86.33%**
Yes	1.04	2.96	**74.00%**	1.04	2.96	**74.00%**	1.04	2.96	**74.00%**

Table 11
Results for DOOP validation for query 2, experiment 4.

'Cookie'	0.5			0.75			1		
	Total	Diff	%	Total	Diff	%	Total	Diff	%
No	360	78	**21.67%**	415	91	**21.93%**	465	110	**23.66%**
Yes	1689	550	**32.56%**	1563	497	**31.80%**	1689	550	**32.56%**

in Experiment 1, a simple experiment was conducted where contextual keywords were used to further limit the types of sentences that were only associated with 'third-party', 'disclose' and 'sharing'. Only sentences with at least one of those keywords mentioned were chosen. The results of this Experiment are reported in Table 12 and Figure 9.

8. Conclusion and Future Work

Privacy policies play an important part in informing users about their privacy concerns. As the world becomes more interconnected and online, security threats prompt users to become more privacy aware, making online privacy policies the primary documents for users making informed decisions. These policies are long and difficult for most users to understand and are infrequently read, presenting a challenge for users. Previous attempts at creating machine readable policies have had limited success as it places the onerous task of crafting these polices on businesses. This paper proposed a novel approach to reducing the amount of text a user has to read by using a domain ontology and NLP to identify key areas of the policies that the user should read to address their concerns and take appropriate action. The approach consisted of constructing DOOP, a domain ontology for online privacy policies, validated against CMU's OPP-115 data set of annotated policies by domain experts. DOOP resulted in 69%, 99%, and 96% reductions in reading for the 3 sample questions, and on average it would take about 45 seconds read the relevant sentences (11 on average). By comparison, the average time to read privacy policies is estimated to be $8 - 12$ minutes [10]. Furthermore, the vocabulary was mapped to the queries stored in the ontology. This allows ontology developers to propose additional insight into related queries and their associated vocabulary. The development of DOOP showed the usefulness of domain ontologies when applied to privacy policies, and also demonstrated a cost-effective way of maintaining and expanding it in the future.

References

[1] M. Madden and L. Rainie, Americans' Attitudes About Privacy, Security and Surveillance, Pew Reseach Center, 2015, Accessed June 10, 2016. http://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance/.

[2] C. Jensen and C. Potts, Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices, *2004 Conference on Human Factors in Computing Systems* **6**(1) (2004), 471–478. ISBN 1581137028. doi:10.1145/985692.985752.

[3] Ten steps to develop a multilayered privacy notice, The Center for Information Policy Leadership at Hunton & Williams LLP, 2006, Accessed June 10, 2016. https://www.huntonprivacyblog.com/wp-content/uploads/sites/18/2012/07/Centre-10-Steps-to-Multilayered-Privacy-Notice.pdf.

[4] M. Munur, S. Branam and M. Mrkobrad, Best Practices in Drafting Plain-Language and Layered Privacy Policies, The International Association of Privacy Professionals, 2012, Accessed June 10, 2016. https://iapp.org/news/a/2012-09-13-best-practices-in-drafting-plain-language-and-layered-privacy/.

[5] A.F. Westin, Privacy and Freedom, Atheneum, *New York* (1967), 7.

[6] Privacy Online: Fair Information Practices in the Electronic Marketplace, Federal Trade Commission, 2000, Accessed June 10, 2016. http://1.usa.gov/1XeBiuY.

[7] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), European Parlaiment, 2016, Accessed June 10, 2016. http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN.

[8] Personal Information Protection and Electronic Documents Act, Senate and the House of Commons of Canada, 2000, Accessed June 10, 2016. http://laws-lois.justice.gc.ca/eng/acts/P-8.6/index.html.

[9] Digital Privacy Act, Senate and the House of Commons of Canada, 2015, Accessed June 10, 2016. http://laws-lois.justice.gc.ca/eng/annualstatutes/2015_32/page-1.html.

[10] A.M. McDonald and L.F. Cranor, The Cost of reading privacy policies, *ISJLP* **4** (2008), 543.

[11] A.M. Mcdonald, R.W. Reeder, P.G. Kelley and L.F. Cranor, A comparative study of online privacy policies and formats, in: *PETS '09: Proceedings of the 9th International Symposium on Privacy Enhancing Technologies*, I. Goldberg and M.J. Atallah, eds, Springer Berlin Heidelberg, 2009, pp. 37–55. ISBN 978-3-642-03168-7. doi:10.1007/978-3-642-03168-7_3.

[12] G.R. Milne and M.J. Culnan, Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices, *Journal of Interactive Marketing* **18**(3) (2004), 15–29, ISSN 10949968. ISBN 10949968. doi:10.1002/dir.20009.

[13] S. Wilson, F. Schaub, A.A. Dara, F. Liu, S. Cherivirala, P.G. Leon, M.S. Andersen, S. Zimmeck, K.M. Sathyendra, N.C. Russell, T.B. Norton, E. Hovy, J. Reidenberg and N. Sadeh, The creation and analysis of a website privacy policy corpus, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, 2016, pp. 1330–1340.

[14] L.F. Cranor, M. Langheinrich and M. Marchiori, A P3P preference exchange language 1.0 (APPEL1. 0), *W3C working draft* **15** (2002).

[15] L.F. Cranor, P3P: Making privacy policies more useful, *IEEE Security and Privacy* **1**(6) (2003), 50–55, ISSN 15407993. doi:10.1109/MSECP.2003.1253568.

[16] R. Lämmel and E. Pek, Understanding privacy policies, *Empirical Software Engineering* **18**(2) (2013), 310–374, ISSN 1382-3256. ISBN 1066401292. doi:10.1007/s10664-012-9204-1. http://link.springer.com/10.1007/s10664-012-9204-1.

[17] Terms of Service; Didn't Read (ToS;DR), Accessed July 12, 2016. https://tosdr.org/index.html.

Table 12

Results for DOOP validation Query 1, Experiment 2 with contextual keywords.

'Contextual Keywords'	**0.5**			**0.75**			**1**		
	TS	**SS**	**Coverage**	**TS**	**SS**	**Coverage**	**TS**	**SS**	**Coverage**
No	91.7	27.99	**30.99%**	91.7	27.99	**30.99%**	91.7	27.99	**30.99%**
Yes	91.7	5.12	**5.57%**	91.7	5.12	**5.57%**	91.7	5.12	**5.57%**

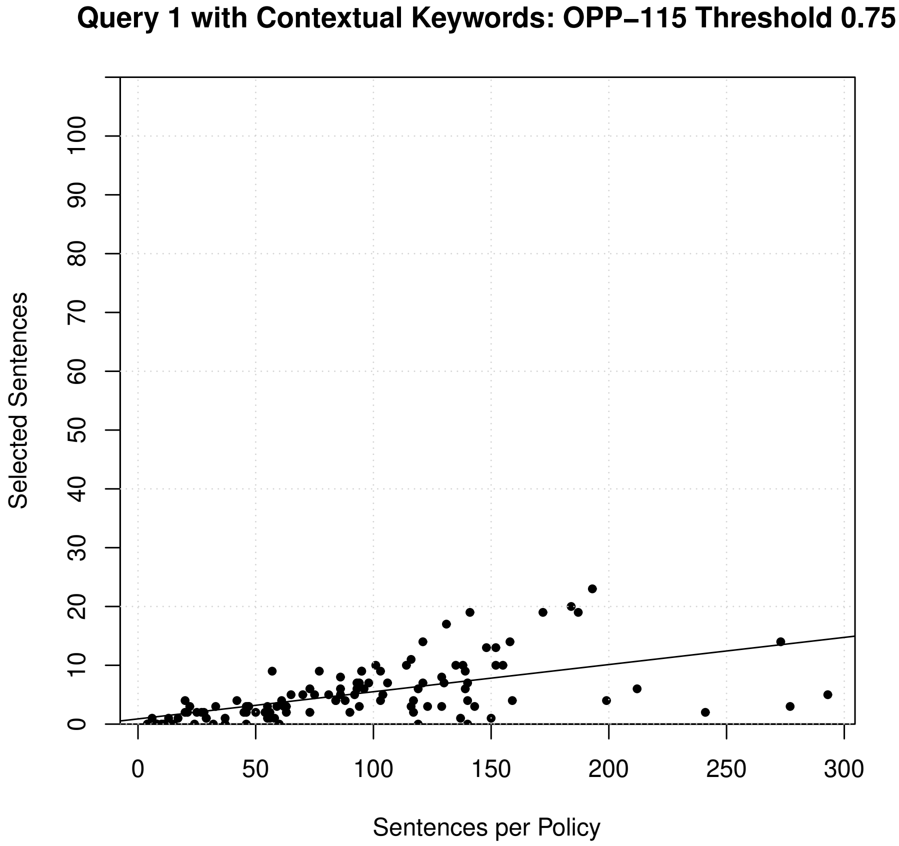


Fig. 9. Results for experiment 2 for query 1 after considering contextual keywords.

[18] S. Zimmeck and S.M. Bellovin, Privee: An architecture for automatically analyzing web privacy policies, in: *23rd USENIX Security Symposium (USENIX Security 14)*, 2014, pp. 1–16.

[19] N. Sadeh, R. Acquisti, T.D. Breaux, L.F. Cranor, A.M. Mcdonalda, J.R. Reidenbergb, N.A. Smith, F. Liu, N.C. Russellb, F. Schaub and et al., The usable privacy policy project: Combining crowdsourcing, machine learning and natural language processing to semi-automatically answer those privacy questions users care about (2013).

[20] R. Ramanath, F. Schaub, S. Wilson, F. Liu, N. Sadeh and N.A. Smith, Identifying relevant text fragments to help crowdsource privacy policy annotations, in: *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[21] S. Wilson, F. Schaub, R. Ramanath, N. Sadeh, F. Liu, N.A. Smith and F. Liu, Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?, in: *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2016, pp. 133–143.

[22] M. Uschold and M. Gruninger, Ontologies and semantics for seamless connectivity, *ACM SIGMOD Record* **33**(4) (2004), 58–64, ISSN 01635808. ISBN 0163-5808.

doi:10.1145/1041410.1041420. http://portal.acm.org/citation.cfm?doid=1041410.1041420.

[23] J. Bermejo, A simplified guide to create an ontology, *Madrid University* (2007). http://tierra.aslab.upm.es/documents/controlled/ASLAB-R-2007-004.pdf.

[24] Y. Malheiros and F. Freitas, A Method to Develop Description Logic Ontologies Iteratively Based on Competency Questions: an Implementation., in: *ONTOBRAS*, 2013, pp. 142–153.

[25] N.F. Noy, D.L. McGuinness et al., Ontology development 101: A guide to creating your first ontology, Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880, Stanford, CA, 2001.

[26] M.C. Suárez-Figueroa, A. Gómez-Pérez and M. Fernández-López, The NeOn Methodology for Ontology Engineering, in: *Ontology Engineering in a Networked World*, M.C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta and A. Gangemi, eds, Springer Berlin Heidelberg, 2012, pp. 9–34. ISBN 978-3-642-24794-1. doi:10.1007/978-3-642-24794-1_2.

[27] M.A. Musen, The protégé project: a look back and a look forward, *AI matters* **1**(4) (2015), 4–12. doi:10.1145/2757001.2757003.

[28] J. Brank, M. Grobelnik and D. Mladenić, A survey of ontology evaluation techniques, *Proceedings of the Conference on Data Mining and Data Warehouses* (2005), 166–170. doi:10.1.1.101.4788. http://eprints.pascal-network.org/archive/00001198/.

[29] K. Liu, W.R. Hogan and R.S. Crowley, Natural Language Processing methods and systems for biomedical ontology learning, *Journal of Biomedical Informatics* **44**(1) (2011), 163–179, ISSN 15320464. ISBN 1532-0464. doi:10.1016/j.jbi.2010.07.006.

[30] P. Velardi, P. Fabriani and M. Missikoff, Using text processing techniques to automatically enrich a domain ontology, *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001* (2001), 270–284. ISBN 1581133774. doi:10.1145/505168.505194.

[31] M.P. Naik, A Survey on Semantic Document Clustering, in: *Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on*, IEEE, 2015, pp. 1–10.

[32] M. Grüninger and M.S. Fox, Methodology for the Design and Evaluation of Ontologies (1995).

[33] N. Fridman Noy and C.D. Hafner, The State of the Art in Ontology Design, *AI Magazine* **18**(3) (1997), 53–74, ISSN 0738-4602. doi:10.1609/aimag.v18i3.1306.

[34] S.C. for Biomedical Informatics Research, Protégé, 2017, Version = 5.2.0. https://protege.stanford.edu/.

[35] D.A. Audich, R. Dara and B. Nonnecke, Extracting keyword and keyphrase from online privacy policies, in: *2016 Eleventh International Conference on Digital Information Management (ICDIM)*, 2016, pp. 127–132. doi:10.1109/ICDIM.2016.7829792.

[36] N. Guntamukkala, R. Dara and G. Grewal, A Machine-Learning Based Approach for Measuring the Completeness of Online Privacy Policies, in: *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2015, pp. 289–294. doi:10.1109/ICMLA.2015.143.

[37] OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, Accessed: 2017-09-10.

[38] D. Tsarkov and I. Horrocks, FaCT++ description logic reasoner: System description, *Automated reasoning* (2006), 292–297.

[39] B. Glimm, I. Horrocks, B. Motik, G. Stoilos and Z. Wang, HermiT: An OWL 2 Reasoner, *Journal of Automated Reasoning* **53**(3) (2014), 245–269, ISSN 1573-0670. doi:10.1007/s10817-014-9305-1.

[40] E. Sirin, B. Parsia, B.C. Grau, A. Kalyanpur and Y. Katz, Pellet: A practical OWL-DL reasoner, *Web Semantics: Science, Services and Agents on the World Wide Web* **5**(2) (2007), 51–53, ISSN 1570-8268. doi:10.1016/j.websem.2007.03.004. http://www.sciencedirect.com/science/article/pii/S1570826807000169.