PrivOnto: a Semantic Framework for the Analysis of Privacy Policies

Editor(s): Mathieu d'Aquin, Knowledge Media Institute of the Open University, United Kingdom; Sabrina Kirrane, Insight Centre for Data Analytics, NUI Galway, Ireland; Serena Villata, INRIA Sophia Antipolis, France
Solicited review(s): Luca Costabello, Fujitsu Knowledge Engineering and Discovery Research Group in Galway, Ireland; Pompeu Casanovas, Universitat AutÃšnoma de Barcelona;

Alessandro Oltramari [a], Dhivya Piraviperumal [a], Florian Schaub [c], Shomir Wilson [d], Sushain Cherivirala [a], Thomas B. Norton [b], N. Cameron Russell [b], Peter Story [a], Joel Reidenberg [b], Norman Sadeh [a]

[a] *Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*
E-mail: oltramale@gmail.com, sadeh@cs.cmu.edu
[b] *Fordham University School of Law, New York, NY 10023, USA*
[c] *University of Michigan School of Information, 105 S. State St., Ann Arbor, MI 48109, USA*
[d] *University of Cincinnati, College of Engineering and Applied Science, 2901 Woodside Drive, Cincinnati, OH 45221*

Abstract. Privacy policies are intended to inform users about the collection and use of their data by websites, mobile apps and other services or appliances they interact with. This also includes informing users about any choices they might have regarding such data practices. However, few users read these often long privacy policies; and those who do have difficulty understanding them, because they are written in convoluted and ambiguous language. A promising approach to help overcome this situation revolves around semi-automatically annotating policies, using combinations of crowdsourcing, machine learning and natural language processing. In this article, we introduce PrivOnto, a semantic framework to represent annotated privacy policies. PrivOnto relies on an ontology developed to represent issues identified as critical to users and/or legal experts. PrivOnto has been used to analyze a corpus of over 23,000 annotated data practices, extracted from 115 privacy policies of US-based companies. We introduce a collection of 57 SPARQL queries to extract information from the PrivOnto knowledge base, with the dual objective of (1) answering privacy questions of interest to users and (2) supporting researchers and regulators in the analysis of privacy policies at scale. We present an interactive online tool using PrivOnto to help users explore our corpus of 23,000 annotated data practices. Finally, we outline future research and open challenges in using semantic technologies for privacy policy analysis.

Keywords: Privacy policies, privacy technologies, ontology-based data access, SPARQL

1. Introduction

As people interact with an increasing number of technologies during the course of their daily lives it has become impossible for them to keep up with the many different ways in which these technologies collect and use their data. Privacy policies are too long and difficult to read to be useful and few, if any, ever bother to read them [28,33]. Yet studies continue to show that people care about their privacy. This results in a general sense of frustration with many people feeling that they have no or little control over what happens to their data. There is a disconnect between service providers and their consumers: privacy policies are legally binding documents, and their stipulations apply regardless of whether users read them. This disconnect between Internet users and the practices that apply to their data has led to the assessment that the "notice and choice" legal regime of online privacy is ineffective in the status quo [36]. Additionally, policy regulators—who are tasked with assessing privacy practices and enforcing standards—are unable to assess privacy policies at scale.

These shortcomings have prompted our team to develop technology to semi-automatically retrieve salient statements made in privacy policies, model their contents using ontology-based representations, and use semantic web technologies to explore the obtained knowledge structures [51]. The research described in this paper focuses in particular on the modeling and knowledge modeling and elicitation part. This includes reasoning about statements that are explicitly made in policies as well as statements that may be missing, ambiguous or possibly inconsistent. End users can benefit from such reasoning functionality, as it can be used to help them better appreciate the ramifications of a given policy (e.g., a statement indicating that a site can share personally identifiable information can be used to infer that the site's policy provides no guarantee that it will not share the user's email address with third parties). Reasoning functionality can also be used to raise user awareness about issues that a policy does not explicitly address or glosses over (e.g. a site that does not mention whether it collects the user's location or shares it with third parties is a site that does not make any guarantee about such practices and therefore one that could engage in such practices). Reasoning can help operators identify potential compliance violations or inconsistencies in their policies, and help them address these issues. Similar functionality can also help regulators check for compliance at scale (e.g. compliance with regulations such as the Children Online Privacy Protection Act, the California Online Privacy Protection Act, or the EU General Data Protection Directive). It can also be used to compare policies within and across different sectors, look for trends over time and more. One can also envision interfaces that could enable end-users to identify alternative websites or mobile apps (e.g., "I don't like that this site provides no guarantee about the sharing of my location: are there other sites offering the same service that will not be sharing my location with third parties?").

We introduce PrivOnto, a semantic technology (ST) framework to model and reason about privacy practice statements at scale. PrivOnto has been validated on a corpus of over 23,000 privacy policy annotations made publicly available by the Usable Privacy Policy (UPP) project, the project that is also the umbrella under which we developed PrivOnto.[1]

[1]Usable Privacy Policy Project: https://www.usableprivacy.org/

The rest of this article is structured as follows. First, we provide overviews of the Usable Privacy Policy Project in Section 2 and related work in Section 3. In Section 4, we describe an ontology of privacy policies populated with about 23,000 annotations of data practices. In Section 5, we illustrate the analysis of the obtained knowledge base with suitable SPARQL queries, designed to pinpoint relevant patterns of privacy practices in the annotated corpus. In Section 6, we provide examples of the semantic search functionality created using the above mentioned SPARQL queries. Finally, in Section 7, we conclude the paper with a discussion of open challenges and directions for future research.

2. The Usable Privacy Policy Project

The Usable Privacy Policy Project builds on recent advances in natural language processing (NLP), privacy preference modeling, crowdsourcing, and privacy interface design to develop a practical framework that uses websites' existing natural language privacy policies to empower users to more meaningfully control their privacy. Figure 1 provides an overview of the approach. We discuss our main research areas below:[2]

Semi-Automated Data Practice Extraction: We aim to extract relevant data practices from privacy policy text in a hybrid approach that combines crowdsourcing and NLP. We leverage crowdsourcing to obtain annotations of privacy policies in terms of topics such as the information collected by a website, whether that information is shared with third parties with or without the user's consent, and whether the collected data can be deleted by users [45]. In parallel, we have developed a corpus of privacy policies annotated by skilled workers with fine-grained detail about the data practices they contain [46]. We plan to use the data from this fine-grained corpus to decompose the annotation task into those subtasks that can be fully automated, such as identification of paragraph topics [48] and user options [49], and those which remain most suitable for crowdworkers.

Privacy Policy Analysis: We use salient information extracted from privacy policies to reason about a website's data practices and conduct extensive privacy policy analysis for multiple purposes. Translating policy features into descriptive logic statements facilitates detection of inconsistencies and contradic-

[2]See [51] for a more complete overview of the project.

tions in privacy policies [50], and annotation disagreement among crowdworkers further helps identifying potential ambiguities in the policy. Comparing a website's privacy policy with those from similar websites holds the potential to detect likely omissions in the privacy policy. Temporal monitoring of changes in privacy policies facilitates content-based trend analysis. Automated analysis of privacy policies and application code can further help identify potential privacy compliance violations, for instance in the context of mobile apps [47]. We use policy analysis results to provide more effective and accurate privacy notices to users. In addition, we plan to make analysis results available to website operators in order to help them improve their privacy policies.

Privacy Preference Modeling: The major goal of our approach is to make privacy policies more usable and accessible for website users. Thus, an important aspect of our work is the identification of those key features in privacy policies that are relevant to users. For this purpose, we have been conducting numerous user studies on privacy concerns, perceptions, and preferences. Furthermore, we strive to gain a deeper understanding of cognitive biases that may negatively affect individuals' privacy decisions, in order to learn how users can be made aware of privacy risks in an effective manner [52].

Effective Privacy User Interfaces: Features extracted from privacy policies as well as results from privacy policy analysis and privacy preference modeling inform our design of user interfaces for privacy notices. The goal is to make those policy features that users care about more accessible, for instance, with nutrition label-inspired privacy notices [53] or privacy icons symbolizing data practices. We are also investigating the potential of just-in-time notices that highlight data practices when they become relevant for the individual user. For instance, data practices concerning the collection and sharing of contact or financial information may only be relevant when the user creates an account or makes a purchase. We are in the process of designing browser extensions that leverage policy extraction results and offer notices to users independently of website operators. We follow a user-centric iterative design process to enhance and evaluate the effectiveness of developed privacy interfaces in user studies.

Finally, in contrast to related work described in the next section, our outlined approach does not require any effort or cooperation by website operators. By making the content of privacy policies more salient and accessible, we hope to also nudge companies towards improving how they present their privacy practices.

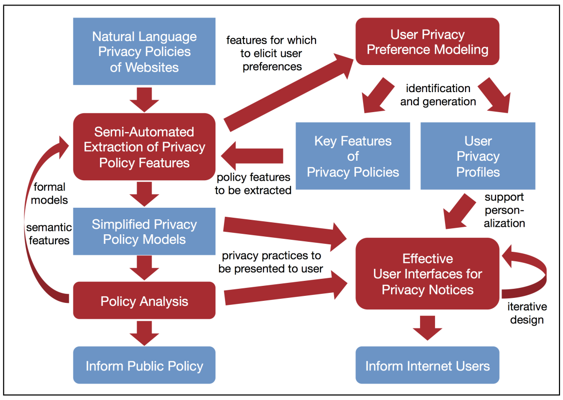

Fig. 1. Overview of the the Usable Privacy Policy Project.

3. Related Work

Privacy-enhancing technologies (PETs) can be defined as the ensemble of technical solutions that preserve the privacy of individuals in their interactions with technological systems. In a recent overview, Heurix et al. [18] categorize PETs along relevant dimensions of privacy, such as the types of data being processed or communicated, application scenarios, grounding in security models, presence of a trusted third party, etc. What their classification fails to account for, however, is the *knowledge dimension* in PETs: without empowering users with the adequate resources to better understand data collection, use and sharing practices, their privacy awareness—the first barrier against any kind of violation—is hindered. In this regard, STs can be considered as knowledge-enabling solutions for PETs, and as support tools for developing context-aware applications [16,43,23,42].

According to Cuenca Grau [15], to be used as effective privacy-preserving systems STs need to embody the following functionalities: (F1) *policy representation*, namely a declarative representation of policies in a system; (F2) *models of interaction*, i.e., a set of queries that can extract relevant information from the system; and (F3) *policy violation*, which formalizes the cases when user preferences and data practices collide, leading to consequences that put users' data at risk. These interconnected functionalities can emerge only when system development follows certain design stages, characterized by Cuenca Grau as: identification of clear privacy requirements and translation into a suitable formal language; realization of the formalized requirements in a computational system; and analysis and verification of the instantiated requirements [26].

PrivOnto, the semantic framework we propose, strives to realize all three functionalities described above, adhering to the related design stages. To the best of our knowledge, most of the existing work on leveraging STs as PETs focuses on defining formal languages for privacy policy representation. For instance, Duma et al. [11] and De Coi & Olmedilla [9] have compared policy languages on the basis of theoretical (e.g., language expressiveness) and empirical principles (action execution, extensibility, etc.). More recently, Bartolini et al. [1] created a legal domain ontology for data protection and privacy, and Breaux et al. proposed 'Eddy' [3], a description logic designed to model privacy requirements, comparing it with alternative – yet less articulated – proposals like KAoS [44], ExPDT [38] and Rein [24]. Eddy has been used to detect conflicts in the specifications of privacy policies, but not yet at large scale. Formalizing policies in the context of description logics was also a goal of the MyCampus and 'PeopleFinder' projects [39,16], which used a semantic web environment in which policies are expressed using a rule extension of the OWL language to capture privacy preferences such as conditions under which users are willing to share their location or other contextual attributes with different services and other users. Other proposals for privacy specification languages include P3P [7], XACML [27], and EPAL [12], though these languages lack formal semantics. A different perspective is taken by Gharib et al. in [54], which presents a new meta-model of privacy ontology, based on a detailed review of the state of the art in privacy requirements engineering.

Policy languages, meta-models and domain ontologies are necessary to implement (F1) and (F3), but are not sufficient to realize (F2). Enabling (F2), namely identifying suitable queries to extract privacy information, is a data-intensive task. In the UPP project we address this issue with an extensive data annotation effort conducted by domain experts. The centrality of (F2) is recognized by Kagal et al. [24] when outlining Rein. Rein is a semantic web framework for representing and reasoning over policies in domains that use different policy languages and knowledge expressed in OWL and RDF-S. Rein realizes a basic version of (F2): a rule-based inference engine checks for relations between a *requester*, a *resource* and some *access properties*. If a relation holds, the output will state whether the *request* is either *valid* or *invalid*. Kagal et al. note that to enhance the privacy and security of web applications more complex, yet user-friendly, query mechanisms need to be implemented. In the next sections, we articulate how this objective is being accomplished in our work by outlining ***PrivOnto***'s architecture and core features. We illustrate how this semantic web framework can be used to model relevant data practices described in natural language privacy policies and augment context-awareness accordingly. We further discuss how ***PrivOnto*** can support privacy engineers and regulators in policy analysis, and provide functionality to also support user-oriented interfaces.

4. *PrivOnto*: Knowledge Base of Privacy Policies

The ***PrivOnto*** knowledge base is comprised of 913,544 RDF triples, obtained by populating a suitable domain ontology with 23,000 annotated data practices from a corpus of 115 privacy policies from US-based companies [46]. ***PrivOnto*** merges a *bottom-up* and a *top-down* approach for ontology creation [30,41]: the former is illustrated in Section 4.1, where we describe the main categories and attributes identified by domain experts to capture data practices expressed in privacy policies; the latter is presented in Section 4.2, where we show how those conceptual structures are formalized as a domain ontology, which has been subsequently populated with a corpus of about 23,000 annotations of data practices. The corpus is described in Section 4.3.

4.1. Domain Expert Frame Analysis of Privacy Policies

In order to study which data practices are expressed in privacy policies, and how data practices are described in privacy policy text, some of the authors and other members of the Usable Privacy Policy Project conducted an iterative multi-disciplinary analysis of privacy policies. The researchers involved in this activity were domain experts with backgrounds in privacy, public policy and law.

4.1.1. Analysis approach

The researchers studied multiple privacy policies of websites from US-based companies drawn from different categories (e.g., news, entertainment, government, shopping) in a iterative qualitative content analysis process. The analysis focused on US websites exclusively. This ensured that the same legal baseline applied to the privacy policy texts and that variations in language would not be attributable to different national legal rules. For example, European law has specific obligations for data practices and notice disclosures that are not found in US law. This means that EU corporate policies would not be accurately compared to US policies based solely on the text's language.

The domain experts would initially read privacy policies individually and mark the types of data practices described in each paragraph of the policy document. Identified types of data practices were then discussed among the researchers and consolidated into consistent codes corresponding to data practice categories. Additional privacy policies were analyzed until no further data practice categories could be identified. This consolidation process was informed by the existing privacy and data protection framework in the United States, including the Federal Trade Commission's Fair Information Practices [14]; the Platform for Privacy Preferences (P3P) [7]; specific privacy notice requirements prescribed by legislation, such as notice requirements in CalOPPA [31], COPPA [6], and the HIPAA Privacy Rule [19]; as well as prior research on privacy policy analysis [35,4,21,8,22]. The combination of content analysis grounded in privacy policy text with the consideration of US privacy legislation and literature ensured that resulting data practice categories are consistent with both (1) how data practices are expressed in privacy policies and (2) the terminology and notice requirements stipulated in US law and literature.

For each of the identified data practice categories, the experts further identified descriptive attributes that collectively represent and define a data practice. For example, a practice describing data collection by the first party (i.e., the website) is defined by how and where information is collected, the type of information being collected and whether it is personally-identifiable information, for what purpose the information is collected, from what user groups information is collected, whether the information is provided explicitly by a user or collected implicitly, and whether users have any choice regarding the practice (e.g., whether they can opt-out). The attributes used to represent data practices, as well as common attribute values were identified in a similar iterative process as the categories, combing the qualitative analysis of attribute and attribute value representations in privacy policy documents with legal requirements in the United States.

This analysis process resulted in a collection of frames that codify the different data practice categories, their descriptive attributes, and typical attribute values as they are expressed in privacy policies. Each frame has its own respective structure of frame-roles and values [13]. These frames were refined over multiple iterations involving their application to additional privacy policies and extensive discussions among the domain experts.

4.1.2. Resulting collection of data practice frames

The resulting collection of frames represents ten categories of data practices, which are defined as follows:

First Party Collection/Use: Privacy practice describing data collection or data use by the service provider operating the service, website or mobile app a privacy policy applies to.

Third Party Sharing/Collection: Privacy practice describing data sharing with third parties or data collection by third parties. A third party is a company or organization other than the first party service provider operating the service, website or mobile app.

User Choice/Control: A practice describing general choices and control options available to users.

User Access, Edit, & Deletion: A practice describing if and how users may access, edit or delete the data that the service provider has about them.

Data Retention: A practice specifying the period and purposes for which collected user information is retained.

Data Security: A practice describing how user data is secured and protected, e.g., from confidentiality, integrity, or availability breaches.

Policy Change: A practice on whether and how the service provider informs users about changes to the privacy policy, including any choices offered to users.

Do Not Track: A practice specifying if and how Do Not Track signals (DNT)[3] for on-line tracking and advertising are honored.

International & Specific Audiences: A Practice that pertains only to a specific group of users, e.g., children, California residents, or Europeans.

Other: Additional sub-labels for introductory or general text in the privacy policy, contact information, and practices not covered by other categories.

[3] https://www.w3.org/2011/tracking-protection/ (W3C Tracking Protection Working Group)

A *data practice* statement belongs to one of these categories, and is characterized by a category-specific set of *attributes*. The frames define a set of potential values for each attribute. Each attribute is supported by a text *fragment* in the privacy policy, which serves as the natural language evidence for the annotated attribute value.

For example, a *First Party Collection/Use* practice is represented by four mandatory and five optional attributes. The mandatory attributes are whether the practice is a positive or negated statement (*Does* or *DoesNot*), how the first party obtained information (*action-first-party*), what kind of information is collected (*personal-information-type*), and for what purpose (*purpose*). In addition, a first party practice statement may indicate whether information is collected implicitly or if the user explicitly provides information (*collection-mode*), whether collected information is linkable to a user's identity (*identifiability*), whether the practice applies to registered users only (*user-type*), and if a user choice is offered explicitly for this practice (*choice-type* and *choice-scope*). Data practices in other categories are represented with similar sets of attributes.

Mandatory and optional attributes reflect the level of specificity with which a specific data practice is typically described in privacy policies. Optional attributes are less common, while mandatory attributes are essential to a data practice. However, the experts' analysis of privacy policies found that descriptions of data practices in privacy policies are often ambiguous on many of these attributes [34]. Therefore, a valid value for each attribute is *Unspecified* in order to express and capture the absence of information. For instance, the fragment "we disclose information to third parties only in aggregate or de-identified form" exemplifies vagueness in data practices as it remains unspecified what information might be disclosed or for what purposes.

This collection of data practice frames constitutes the semantic foundation for the ***PrivOnto*** ontology, described in the next section.

4.2. Domain Ontology for Privacy Policies

The ***PrivOnto*** ontology is a formal model of the data practices identified by domain experts. It represents unstructured policy contents according to frame-based structures specified using OWL-DL. In ***PrivOnto***, each data practice category is modeled as a class characterized by a wide spectrum of Object and Datatype properties (see Figure 2): we used the latter to represent the specific attributes of each category, which essentially correspond to the backbone of the collection of frames presented in the previous section; conversely, the former were used to represent the conceptualization of the domain, and delineate the semantic relations holding between the defined classes.

The Object property `denote` holds between the class $ANNOTATION$ and the class $SEGMENT$: the resulting pattern captures the difference between annotations, namely the entities that emerge from tagging discrete parts of privacy policies with suitable frames and roles, and the specific text they refer to. Accordingly, individual annotations `denote` individual segments (policy paragraphs) and their constituent parts or *fragments*. The class $SEGMENT$ and the class $FRAGMENT$ are linked by the `part_of` relation, which is axiomatized as asymmetric and irreflexive. This semantic structure reflects the compositionality of paragraph-length segments: fragments can span from single words to well-formed sentences, whereas segments correspond to syntactically and semantically coherent sequences of fragments. By means of the `part_of` relation, the same segment can instantiate multiple data practices via its fragments.

Fragments are labeled with a unique identifier (UID), consisting of the policy number, the segment number, and the start and end indexes of the selected text. In the same way, we assigned UIDs to instances of practice categories. Thanks to this modeling strategy, we can refer to different annotations of the same fragment, so that the "raw" policy content is kept distinct from all the annotations that refer to it. For example, a fragment stating that "by use of our websites and games that have advertising, you signify your assent to SCEA's privacy policy" is annotated as an instance of **First Party Collection** and as an instance of **User Choice**, reflecting different aspects of the policy text. This situation can be represented in ***PrivOnto*** by two instances of $ANNOTATION$, each exemplifying different data practice categories, and referring to the same individual of $FRAGMENT$. The actual content of a fragment is expressed in the form of 'string' values in the range of the `annotated_text` datatype property, whose domain is the $FRAGMENT$ class. For example fragment `3819-3-95-203` is associated with the following statement "The information we learn from customers helps us personalize and continually improve your Amazon experience." This fragment is used in Figure 3, which shows how annotations, data practice categories and fragments are connected in the

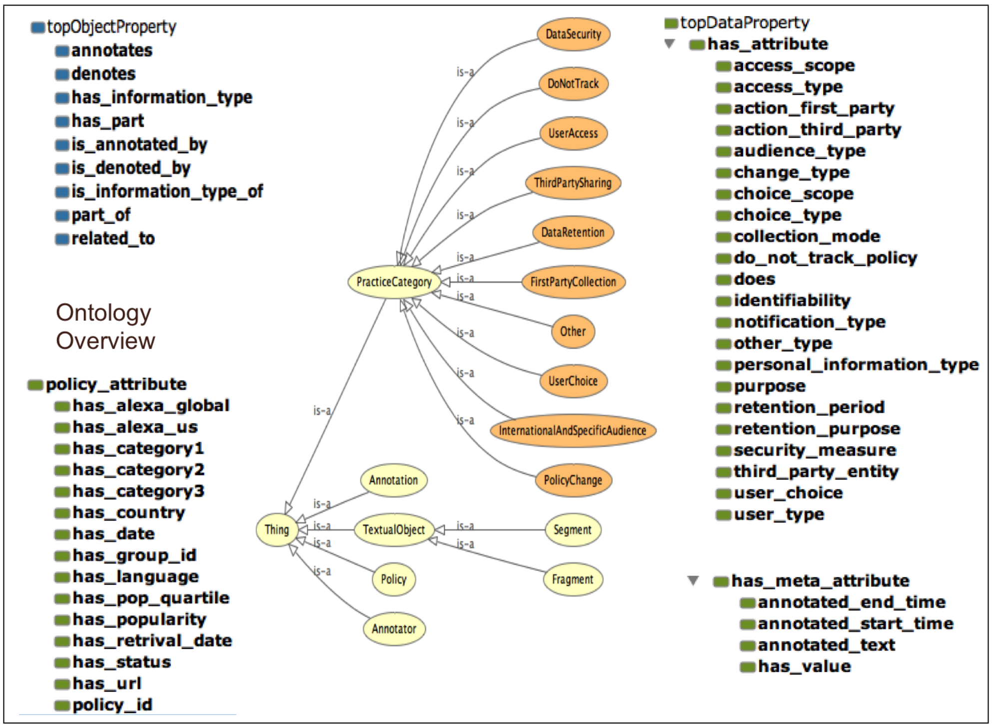

Fig. 2. Protégé visualization of ***PrivOnto*** hierarchies of Classes, Object properties and Datatype Properties.

ontology. The ***PrivOnto*** framework does not directly address the linguistic structures of a given policy, but it pinpoints them only insofar as they instantiate a data practice category: we demonstrate in Section 5 how this is actually a key strength of our approach.

The ontology also includes $ANNOTATOR$, a class whose instances denote the individuals involved in the annotation task: the relation `executed_by` between $ANNOTATION$ and $ANNOTATOR$ preserves the traceability of the identified data practices.

PrivOnto also includes general information about the website where the privacy policy can be found: the date when it was crawled, contact information of the company to which the policy belongs, the company's website, the associated Alexa's traffic ranking information,[4] etc. Note that some of this 'meta-information' is subject to change, and thus needs to be regularly monitored and documented: to this end, ***PrivOnto*** supports `xsd:dateTime` values, which serve as temporal indexes for policies' meta-information. Privacy policies may vary over time as well: in this case it is not only important to record changes, but also to investigate their implications: policies are systematically updated by companies for a variety of reasons, and analyzing the consequences of these modifications to enforced data practices is of key importance to regulators and users. The privacy policies obtained for annotation were collected at the same time, thus policy changes do not occur in our dataset. Nevertheless, future expansion of our corpus will include the addition of new privacy policies along with updates to already represented policies. We therefore plan to extend ***PrivOnto*** with `OWL-Time`[5] to enable qualitative and quantitative temporal reasoning [20].

4.3. *Corpus of Annotated Privacy Policies*

PrivOnto was instantiated based on the OPP-115 corpus [46], a corpus of 115 privacy policies of US-based companies, each independently annotated by three legal experts according to the developed collection of data practice frames. In this section, we characterize the OPP-115 corpus and the annotation process.

Privacy policies vary along many dimensions of analysis, including length, legal sophistication, readability, coverage of services, and update frequency.

[4] http://www.alexa.com/topsites/countries/US

[5] https://www.w3.org/2001/sw/BestPractices/OEP/Time-Ontology-20060518

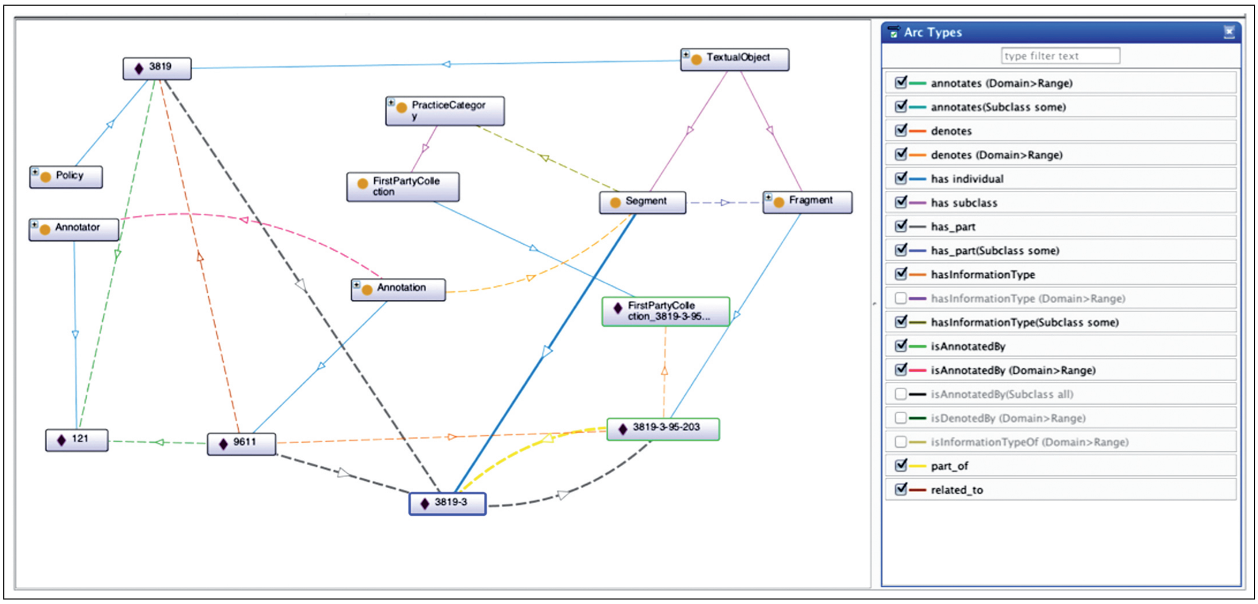

Fig. 3. LEFT: an example that shows how ***PrivOnto*** structures are used to model the semantic relations between data practices, fragments and segments of policies. RIGHT: legenda of semantic relations (redundant arcs are grayed-out to simplify the figure).

Large companies' policies may cover multiple apps, services, websites, and retail outlets, while privacy policies of smaller companies may have narrower scope. Accordingly, privacy policies were chosen for inclusion in the UPP corpus using a procedure that encouraged diversity.

Websites were selected using a two-stage process: (1) relevance-based website pre-selection and (2) sector-based subsampling. This first stage consisted of monitoring Google Trends [17] for one month (May 2015) to collect the top five search queries for each trend; then, for each query, the first five websites were retrieved on each of the first ten pages of search results. This produced a selection of 1,799 unique websites. For the second stage, websites were chosen from each of DMOZ.org's top-level website sectors (e.g., News, Shopping, Arts).[6] Note that the DMOZ.org's "World" sector was excluded and that the "Regional" sector was limited to the "U.S." subsector in order to exclude non-US privacy policies and to insure that all policies were subject to the same legal baseline.

For each sector, eight websites were selected based on occurrence frequency in Google search results. More specifically, the eight websites were randomly selected two-apiece from each rank quartile. Each selected website was manually verified to have an English-language privacy policy and to belong to a US company (according to contact information and the website's WHOIS entry). Websites that did not meet these requirements were replaced with random redraws from the same sector and rank quartile. Notably, some privacy policies covered more than one selected website (e.g., the Disney privacy policy covered disney.go.com and espn.go.com). The consolidation of the corpus resulted in a final dataset of 115 privacy policies of US-based companies across 15 sectors.

We developed a web-based annotation tool, shown in Figure 4, to facilitate annotation of the UPP corpus' privacy policies by expert annotators according to our frame-based annotation scheme. Privacy policies were divided into *segments* and shown to annotators sequentially in the tool. Each segment may be annotated with zero or more data practices from each category. To annotate a segment with a data practice, an annotator assigns a practice category and specifies values and respective text spans (*fragments*) as appropriate for each of its attributes.

Each privacy policy was independently annotated by three expert annotators. In total, we hired 10 law students as experts on an hourly basis to annotate the complete set of 115 privacy policies. Note that the average annotation time per policy was 72 minutes. The annotation of the corpus resulted in about 23,000 annotations of data practices, which were used to popu-

[6]The DMOZ.org website sectors are notable for their use by Alexa.com.

late the ***PrivOnto*** ontology and create the corresponding knowledge base.

5. Query-based Semantic Analysis of Privacy Policies

PrivOnto facilitates the elicitation of prominent information from privacy policies in order to gain insights on the nature of data practices. This knowledge elicitation process leverages a library of 57 SPARQL queries[7] we engineered to retrieve data practice categories, attributes, and values from the annotated corpus.[8] Our work required only marginal effort for translating unstructured natural language questions into formal queries, as our frame-based annotation process embedded 'saliency' in the corpus of annotations in the form of ontology categories and attributes. For this reason, the ontology-based analysis of privacy policies proposed in this article did not require dealing with the diversity and ambiguity of natural language text [25]. The queries we present in Section 5.2 match *by design* the privacy questions that domain experts deemed as relevant for policy analysis, and that originated the ***PrivOnto*** framework in the first place.

5.1. Architecture

Our architecture for mapping the structured annotation corpus to the ***PrivOnto*** ontology is shown in Figure 5. The mapping process resulted in a .owl file that captures the corpus (913,544 RDF triples). The obtained knowledge base was then loaded in an Apache Jena Fuseki server[9] for dynamic processing: the server provides a web service framework for different applications to access data through SPARQL queries. Figure 6 shows the ***PrivOnto*** semantic web environment. This API was further used by Usable Privacy Policy website to create a semantic search tool for querying privacy policies.

5.2. Library of Queries

We created 57 SPARQL queries to analyze different aspects of the 115 privacy policies represented in the ***PrivOnto*** ontology: this method enabled us to build a scalable semantic retrieval system for gaining insights on privacy practices related to the collection, use, and sharing of personal data. The queries in the library can be categorized by two orthogonal dimensions, based on: (1) the type of targeted information (quantitative, qualitative, truth-values) and (2) the selected practice category.

It is important to point out that all 57 queries return the annotated text associated with a policy fragment: this feature realizes a crucial aspect of *model of interaction* (see functionality F2 in Section 3), i.e., the possibility for legal experts and users to understand and evaluate the machine-readable semantic models and queries in relation to a privacy policy's original text.

Table 1 shows the different kinds of information that can be extracted from the knowledge base, along with sample queries. Percentage and count type questions help gain an overall understanding of the privacy policy data.

For example the query below, which calculates the *'number of policies that allow users to export their data*,' returns 1 as the answer. Thus, only one out of 115 policies in our data set provides for the export of collected data, which shows the exceptionality of this data practice in the considered dataset.

```
SELECT (COUNT(*) AS ?count) {SELECT DISTINCT ?policy
 WHERE {?p a privonto:UserAccess.
         privonto:access_type "Export"^^xsd:string.
         privonto:related_to ?policy.}
```

In order to verify facts in the ontology, we can use ASK queries. For instance, the query below, which matches the question *'Does any policy state that personal information is shared or collected as part of a merger?*,' returns `True` as output. By replacing the ASK clause with a SELECT clause, we can easily assess that nine policies include that data practice.

```
ASK
 WHERE
  {?frag privonto:part_of ?segment.
   ?frag privonto:has_information_type ?practice.
   ?prc  privonto:purpose "Merger/Acq"^^xsd:string.
   ?prc  privonto:related_to ?policy.
   ?prc  a privonto:FirstPartyCollection.}
```

Our SPARQL queries also help gain specific information about different practice categories. For instance, the query exemplified by the question *'How many websites mention each audience type?'* lead us to discover that clauses are generally added for children (86 out of 115 privacy policies), which suggests

[7] Version 1.1: https://www.w3.org/TR/2013/REC-sparql11-query-20130321/

[8] Despite being extensive and detailed, this library is not meant to be exhaustive, and can be further expanded.

[9] https://jena.apache.org/download/index.cgi

Fig. 4. Web-based tool for expert privacy policy annotation.

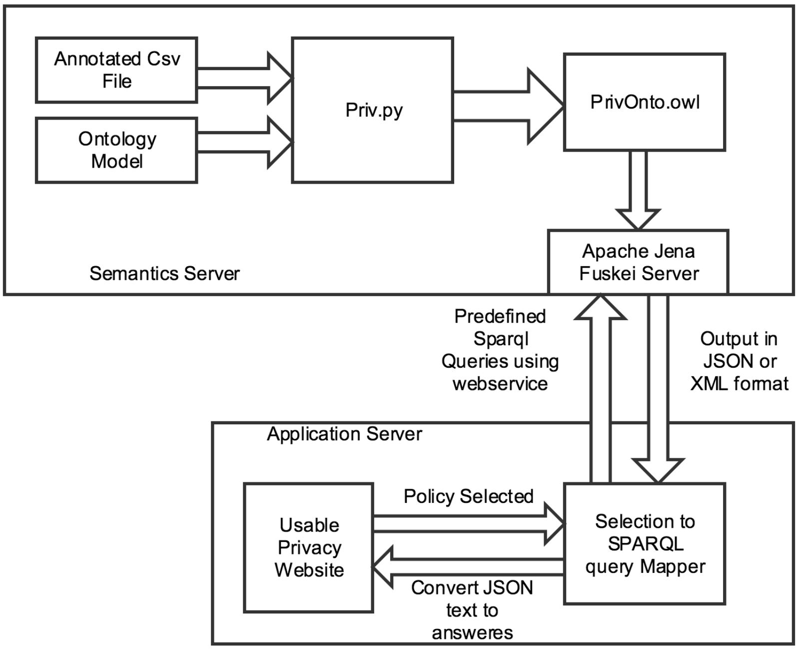

Fig. 5. Semantic server architecture for querying ***PrivOnto***.

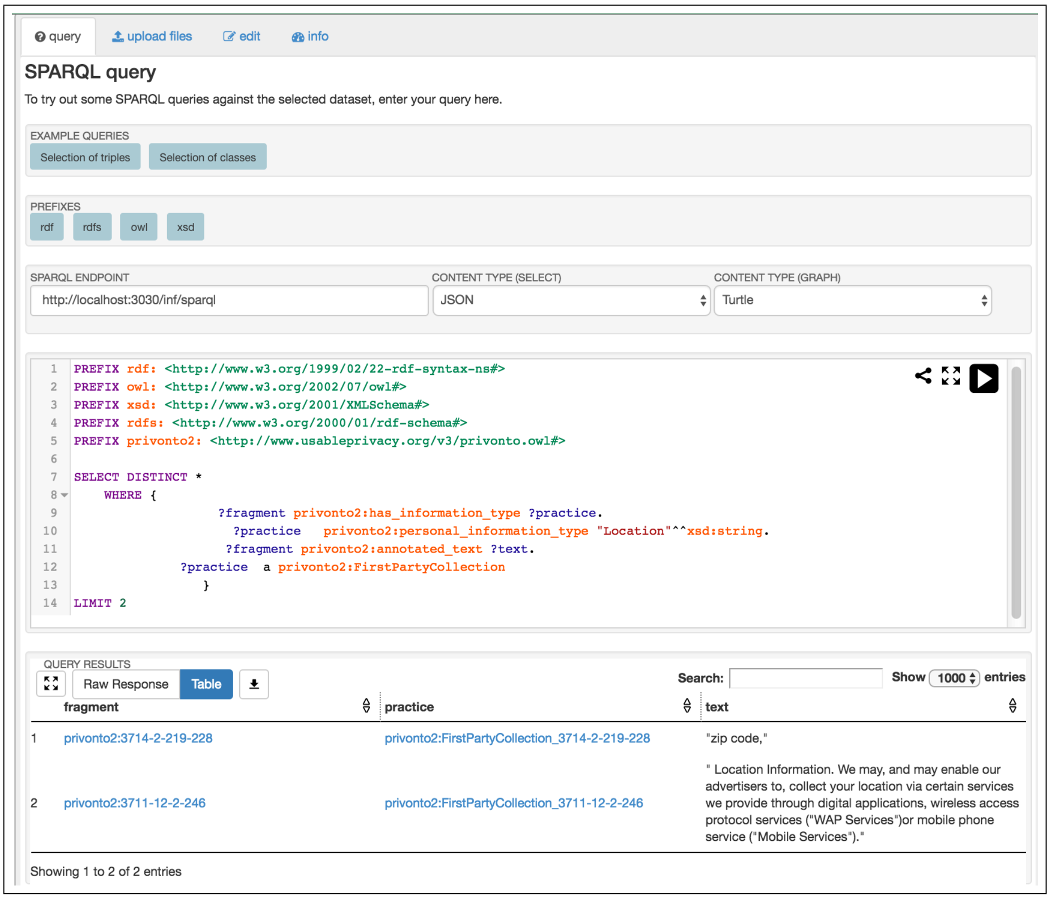

Fig. 6. Screenshot of the Apache Jena Fuseki server used for querying ***PrivOnto***: the query in the example returns two policy fragments about collection of location information. Note that the `LIMIT 2` clause was used to fit the results to the window's size.

that a large number of privacy policies aim to be compliant with the Children Online Privacy Protection Act (COPPA) [6], but also shows that 25% of the privacy polices in our corpus have no provisions specific to children.

The second dimension through which our SPARQL queries can be classified is based on different practice categories. Each practice category provides very specific information about privacy policies. By organizing the queries in this way, we can concentrate on specific characteristics of a policy, and draw parallel conclusions from different categories. Table 2 shows example queries from each category.

While running experiments in the Jena Fuseki environment, we observed that the queries' processing time depends on the complexity of the SPARQL expression, while being only partially correlated with the number of matches. In particular, Figure 7 represents the proportion between number of matches and retrieval times for a subset of 20 SPARQL queries chosen across all data practice categories to highlight relevant types of information in a policy. For instance, the figure shows that only four queries had processing time higher than 1500 ms: these queries included SPARQL constraints like OPTIONAL and MINUS. The queries labeled as 'Financial Information and Purpose', 'General Information and Purpose', 'Unspecified Information and Purpose' refer to user's collected information at different levels of granularity, and specify the purpose of collection only when found in a policy: this condition was expressed in the SPARQL request by an OPTIONAL clause on the 'Purpose' attribute of the 'First Party Collection/Use' category. In the case of the query labeled as 'Policies with User Choice,' the high processing time was brought about by the MINUS clause, introduced to discard from the results all the policies with no real user choice, but only with take-it-or-leave-it option (this aspect is further analyzed in section 5.3.3).

Table 1

Targeted information and related query types.

Targeted Information	Query example
Percentage	What percentage of policies apply to websites and mobile apps?
Count on Practices	How many practice statements per policy are unclear about where information are collected from users?
True or False	Is information shared or collected as part of a merger or acquisition?
Count on Policy	How many policies have statements on user choice?
Count on Supporting text in each Policy	For each of the security-measure values, how many websites mention them?

Table 2

Queries are sent to the Apache Jena Fuseki server that runs the ***PrivOnto*** framework: quantitative results shown in the table indicate the number of fragments, number of policies, and percentages related to specific data practices.

Category	Type of Queries	Result
First Party Collection	Fragments that collect finance information and for what purpose?	231
Third Party Sharing	Fragments that denote user information is shared with external third parties	2,220
User Choice	How many policies have statements on user choice?	106
User Access	Percentage of policies that allow users to delete their account	0.18
Data Retention	Percentage of statements where a period is stated for data retention	0.09
Data Security	For each of the security-measure values, how many websites mention them?	10
Policy Change	How many websites specify a user choice on policy change?	91

5.3. Results

In this section we provide an overview of the quantitative and qualitative results of our query-based semantic analysis of about 23,000 data practices instantiated in the ***PrivOnto*** knowledge base.

5.3.1. Personal information collection/sharing

For the practice categories *User Choice*, *First Party Collection/Use*, and *Third Party Sharing/Collection*, we observed that privacy policies specify the information collected or shared, though the purpose of data collection is rarely mentioned in the same fragment. Therefore, we collected the purpose information from the other fragments present in the parent segment. We observed that, apart from 'unspecified,' 'basic service' and 'additional service' were the most mentioned purposes. 'Device information' and user's 'online activity' are collected from users' for 'analytics/research' purposes, whereas 'finance' and 'contact information' were collected for 'marketing' and 'advertising purposes.' Purpose for which information is highly shared is 'Advertising' (14.6%), and the purpose for which information is highly collected is for 'basic service/feature' (16%).

Table 3 presents the comparison of different personal data types which are collected and shared. We observed that most of the data types collected and shared are unspecified (last row). This result can be explained by the fact that the word "information" is often used with no further description or specification in the policies. As a result, the privacy policies make it difficult for consumers and regulators to determine which information is actually collected or shared by a company. The following text fragments exemplify this vagueness: "the information we learn from customers helps us personalize and continually improve your Amazon experience" and "any information that we collect from or about you."

Table 3 also shows that 'device,' 'location identifiers,' and 'contact information' are often collected by the websites, but are not explicitly mentioned in statements with respect to third party sharing. Because of the extensive use of generic descriptions for information types, the privacy policies do not indicate whether these data items are actually shared with third parties.

'Contact information,' 'user online activities,' and 'general personal information' are the top referenced types of information. 'Contact information' appears

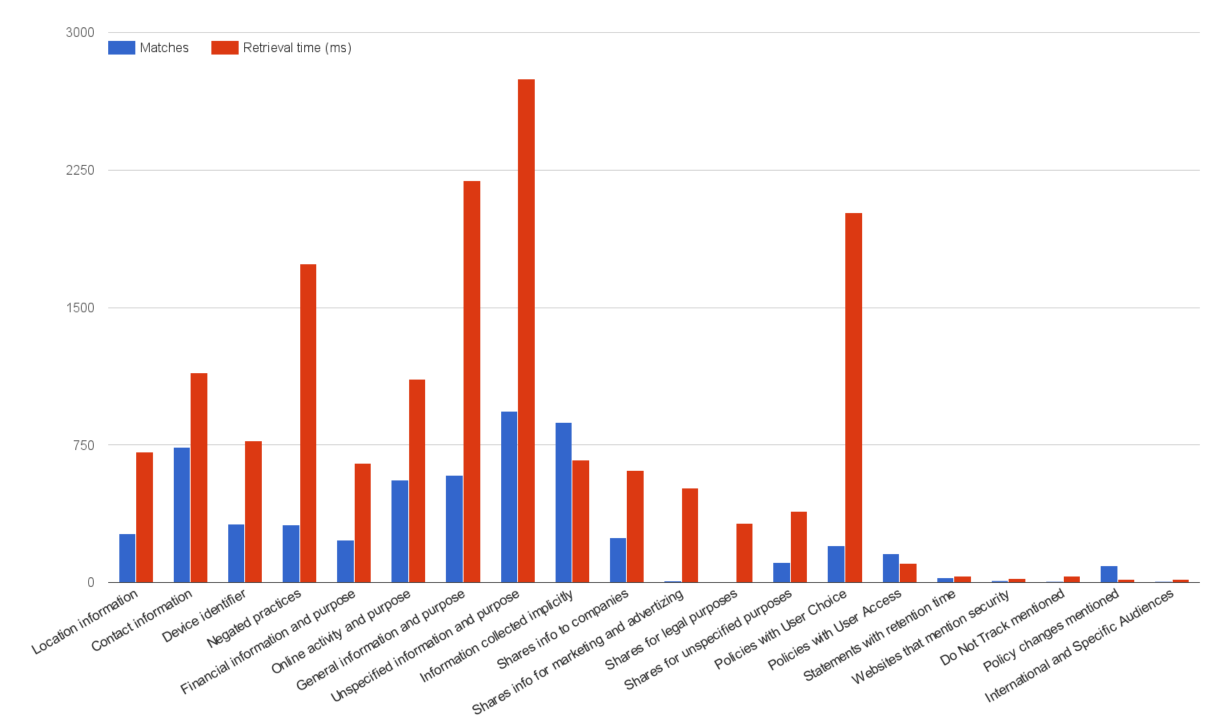

Fig. 7. Proportion between number of matches and processing times for a subset of 20 queries. The labels in the *x-axis* represent types of information collected, shared, or mentioned in a policy and returned by suitable SPARQL queries. The *y-axis* represents the corresponding number of matches (blue histograms) and the retrieval time in milliseconds (red histograms).

Table 3

Queries on information collected from users or shared about users. Number of fragments are visualized, as well as coverage across policies,

Question	First Party Collection	% Policies	Third Party Collection	% Policies
Fragments that collect/share location information and for what purpose?	265	59.13	61	26.09
Fragments that collect/share contact information and for what purpose?	736	90.43	246	57.39
Fragments that collect/share device identifier and for what purpose?	319	76.52	75	25.22
What kind of Fragments are especially negated	199	67.83	313	78.26
Fragments that collect/share finance info and for what purpose?	231	63.48	102	35.65
Fragments that collect/share user's online activities info and for what purpose?	559	87.83	294	66.96
Fragments that collect/share user's general personal information info and for what purpose?	587	88.70	730	91.30
Fragments that collect/share user's unspecified info and for what purpose?	936	85.22	820	88.70

frequently as collected information, while 'general personal information' is highly shared. 'General personal information' is also often ambiguous. The corresponding policy fragments describe this information as "personally identifiable information" or "personal information." For example, one policy in the corpus shares "any and all personal identifiable information collected from our customers" with third parties.

Out of 115 policies, 90 privacy policies state that the service providers do not share some information with third parties, and 78 policies explicitly state what information they do not collect from users. The top categories of information type reportedly not collected or not shared are 'generic personal information,' 'cookies and tracking elements,' and 'contact' information. While this appears to contradict the previous finding that contact information is frequently collected and general personal information is widely shared, the contradiction reflects that privacy policies are explicit when they do not share data.

5.3.2. Marketing and Advertising

There were 886 fragments which described the collection of information for 'Marketing' and 'Advertising' purposes. Information collected for advertising purposes is typically identified as the user's 'online activities' or 'cookies and tracking elements'. Users' 'contact' information is typically used for 'marketing' purposes.' By contrast, 'financial' information is often identified for sharing with third parties when these are partners or affiliates.

5.3.3. User's choice on enabling service

Almost all privacy policies (92%) have statements describing User Choices. But, of these privacy policies, 48% have statements that merely describe a take-it-or-leave-it choice. Instead of a real choice, users are told not to use the service or feature if they disagree with the privacy policy or with certain data practices. Examples are: "if you choose to decline cookies, you may not be able to fully experience the interactive features of this or other Web sites you visit" or "if you do not agree to this privacy policy, you should not use or access any of our sites."

5.3.4. User Data Retention

About half of the privacy policies (56%) specify for how long they store user data. In 40% of these policies a retention period is explicitly 'stated' (e.g., 30 days) or the retention period is at least 'limited' (e.g., stored as long as needed to perform a requested service); while 7% express that the data will be stored indefinitely. The distinction between 'Limited' and 'Stated' retention periods is sometimes blurred due to drafting vagueness and annotator interpretation. For instance, the fragment "we will retain your data for as long as you use the online services and for a reasonable time thereafter" has been annotated both as "limited period" or as "stated period." This creates ambiguity with respect to the duration that user data will remain in a service's database.

5.3.5. Data Export

As mentioned in the previous section, only one policy in our knowledge base describes how users can export data. The respective annotated fragment states: "California Civil Code Section 1798.83, also known as the Shine The Light law, permits our users who are California residents to request and obtain from us once a year, free of charge, information about the personal information (if any) we disclosed to third parties for direct marketing purposes in the preceding calendar year."

5.3.6. Policy Change

Privacy policies typically provide that users are notified about changes to the privacy policy through some form of general notice or through a website. Only 30% of the privacy policies containing descriptions of change in notification practices mention a notification of individual users (e.g., via email). The lack of personal notice for policy changes means that users are unlikely to be aware of changes to the privacy policy, although such changes may alter how information about them is collected, used, or shared by a service.

5.3.7. Data Security

The major security measures which most websites describe are the use of 'secure user authentication,' the existence of a 'privacy/security program,' and the communication of data with 'secure data transfer.'

The analysis above shows that query-based analysis of the ***PrivOnto*** knowledge base can provide insights on privacy policy data both on a semantic and textual level. We can both verify information and collect statistics on privacy policies by means of the ***PrivOnto*** semantic framework. Ontology-driven analysis can help distill the content of a privacy policy, as well as help compare the target policy with similar policies. In this respect, ***PrivOnto*** can help users gain insights on the stated practices of services they use and help them make more informed privacy choices.

6. Semantic Search

In the previous section, we analyzed the knowledge base created using ***PrivOnto*** ontology. While SPARQL is a very useful framework to acquire information from a OWL ontology, it is not easy for a layman to work with. SPARQL expertise is crucial in extracting the correct information from a knowledge base. In order to make our work user friendly, we decided to create a semantic search functionality where natural language queries will be converted to SPARQL queries for easy access. The UPP portal already visually integrates the data practice annotations with a privacy policy's original text in an easy-to-use web interface (see Figure 11), and enables users to filter for attributes and values of specific frame categories, although currently in a limited manner without the support of semantic technologies.

We have extended this functionality as a part of our UPP project's data exploration portal.[10] As shown in Figure 5, natural language queries were mapped to SPARQL queries at the application server end. Using the web API created by Jena, answers to the queries were retrieved from the semantic server. Depending on the type of the queries, qualitative answers where shown as a paginated table and quantitative queries where shown as a interactive bar chart. Figure 8 shows the result for both the type of queries. Currently, the initial version of this search functionality is under beta testing phase in our development server.

In the initial version of the semantic search, we are presenting the users with the natural language queries. They can filter these queries based on the practice categories and question type as discussed in the previous section. For quantitative queries which extracts part of text from a website policies, we provide link to the paragraph of the policy the text comes using a link in website name column. Users can use this link to get more clarity on the results. Figure 9 shows an example of this functionality. For users who are interested in knowing the actual SPARQL query behind the results, a small button is added (see Figure 9), to show the underlying SPARQL query.

7. Discussion and Future work

In this paper we described ***PrivOnto***, a semantic web framework used to represent data practices in privacy policies and support knowledge elicitation. ***PrivOnto*** is an essential tool for regulators and can also enable more usable privacy notices by exposing semantic reasoning results to users. We show the utility of ***PrivOnto*** by instantiating it with a corpus of 115 privacy policies of US-based companies which have been annotated by domain experts as part of the Usable Privacy Policy project.

The ***PrivOnto*** ontology model formalizes a frame-based annotation scheme that helps experts identify data practices in policy text. As a result, each relevant fragment of a policy has been mapped to suitable ontology categories and attributes, generating a knowledge base of about 23,000 annotated data practices. Each fragment may be associated with different categories and attributes, on the basis of interpretations by multiple annotators. In this regard, consolidating alternative and potentially conflicting interpretations is a relevant challenge for our work, which we are currently addressing using natural language processing and machine learning techniques.

To the extent that contradictions have a logical nature, state-of-the-art inference engines like Pellet [32] would be sufficient to flag them. For instance, preliminary results show that there's complete agreement when it comes to annotate if a *Do Not Track* data practice is 'honored' or 'not honored' by a given policy: but in cases when those two mutually exclusive values were to be selected for the same fragment, automatic reasoning with ***PrivOnto*** would detect the inconsistency.

PrivOnto's semantic representation and current knowledge base is grounded in data practice annotations of US companies' privacy policies and framed by US privacy law and standards. The described annotation and modeling process can be replicated to derive knowledge representations for privacy policy content subject to other legal and regulatory frameworks, e.g. in Europe or the Asian-Pacific region. Ontologies associated with other legal or regulatory frameworks could be developed to facilitate compliance analysis across privacy and data protection requirements in different regions and contexts.

Semantically-labeled privacy policies constitute an important resource for privacy analysts and regulators, but scaling the process of annotating natural language privacy policies accordingly can be challenging. As part of the efforts in the UPP project, we investigate the potential of crowdsourcing privacy policy analysis from non-experts, in combination with machine learning, in order to enable semi- or fully automated extrac-

[10]https://explore.usableprivacy.org/

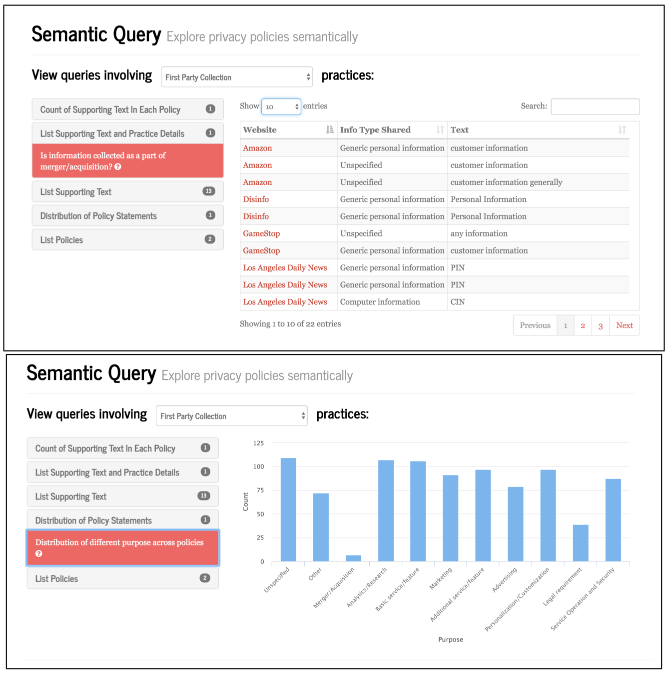

Fig. 8. Two screenshots of the qualitative and quantitative results visualized as a table in semantic search

Fig. 9. A screenshot of the search functionality in UPP website.

tion of data practices and their attributes from privacy policy documents [2,5,45]. These efforts show promise for scaling up our analysis, which would enable further expansion of ***PrivOnto***'s knowledge base.

PrivOnto shows how STs can be used to provide privacy researchers, regulators, site operators and end users with practical reasoning functionality that can help them deal with the complexity of privacy policies. This includes using inferences to highlight important ramifications of privacy policy statements. These inferences can help end users see how some policy statements (or lack thereof) align with their actual concerns (e.g. "could this site possibly share my location with third parties?", "for how long does this site keep my location data?"). They can help site operators identify inconsistencies in their policies (e.g. a site stating that it does not share Personally Identifiable Information (PII), yet indicates that it shares email addresses with third party affiliates). They can help regulators identify potential compliance violations. They

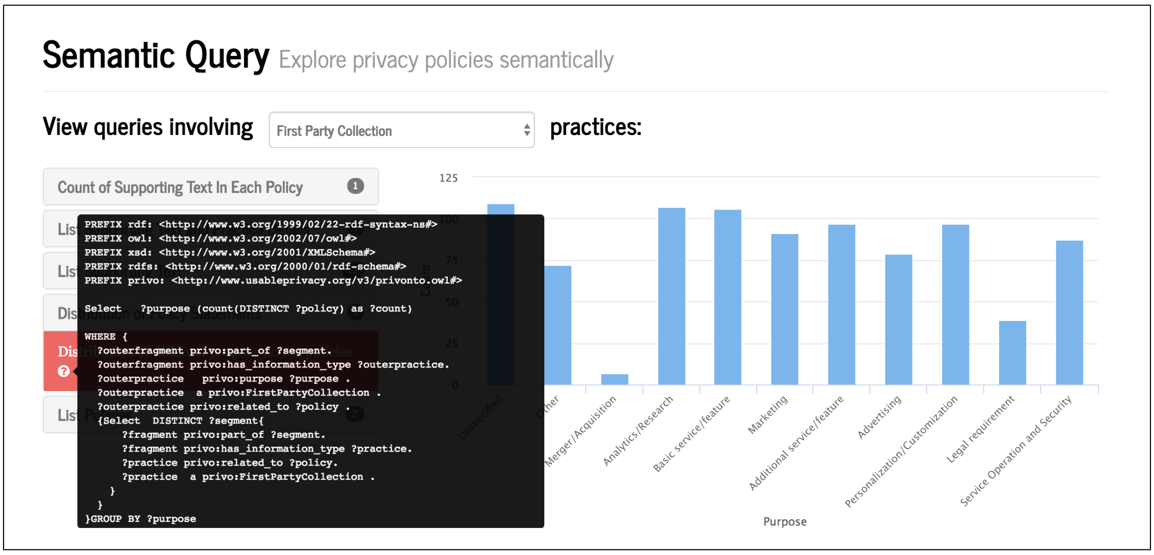

Fig. 10. SPARQL version of a query in the search page

could ultimately also support more sophisticated interfaces that empower users to identify alternative sites or apps similar to the ones they are currently considering but without privacy practices with which they may not feel comfortable. The search functionality presented in this paper revolves around an initial set of 57 SPARQL queries derived from conversations with privacy scholars, including both legal scholars and experts in modeling people's privacy concerns, given our objective of supporting reasoning functionality capable of supporting a broad range of usage scenarios. Over time we envision further refining this set of queries, as we continue to collect feedback from different target user communities (end-users, site operators and regulators). We also envision creating extensions of the framework presented herein, where annotations collected from multiple annotators are combined and assigned confidence levels that reflect the level of agreement among annotators. These confidence levels could in turn be combined according to some logic when assigning confidence levels to facts inferred from consolidated annotations – a number of different possible frameworks are available here.

8. Acknowledgments

This research has been partially funded by the National Science Foundation under grant agreements CNS-1330596 and CNS-1330214. The authors would like to acknowledge the entire Usable Privacy Policy Project team for its dedicated work; and especially thank Pedro Giovanni Leon, Mads Schaarup Andersen, and Aswarth Dara for their contributions to the design and validation of the annotation scheme, as well as the corpus creation.

References

[1] Bartolini, C., Muthuri, R., and Santos C.: Using Ontologies to Model Data Protection Requirements in Workflows. In Intl. Workshop Juris-informatics (2015)

[2] Bhatia, J., Breaux, T.D., and Schaub, F.: Mining Privacy Goals from Privacy Policies using Hybridized Task Re-composition. ACM TOSEM (forthcoming)

[3] Breaux, T.D., Hibshi, H., and Rao, A.: Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements. Requirements Engineering, 19.3, 281-307 (2014)

[4] Breaux, T.D. and Antón, A.: Analyzing Regulatory Rules for Privacy and Security Requirements. IEEE Trans. Software Eng., 34.1, 5-20 (2008)

[5] Breaux, T.D., and Schaub, F.: Scaling requirements extraction to the crowd: Experiments with privacy policies. In Intl. Req. Eng. Conf., IEEE (2014)

[6] Children's Online Privacy Protection Rule (COPPA). 16 CFR Part 312 (1998)

[7] Cranor, L.: Web privacy with P3P. O'Reilly Media, Inc.(2002)

[8] Cranor, L.F., Leon, P.G., and Ur, B.: A Large-Scale Evaluation of U.S. Financial Institutions' Standardized Privacy Notices. ACM Trans. Web, 10.3 (2016)

[9] De Coi, J.L., and Olmedilla, D.: A review of trust management, security and privacy policy languages. In Int. Conf. Security and Cryptography (2008)

[10] De Greene, K. B.: Sociotechnical systems: factors in analysis, design, and management. Prentice-Hall, (1973).

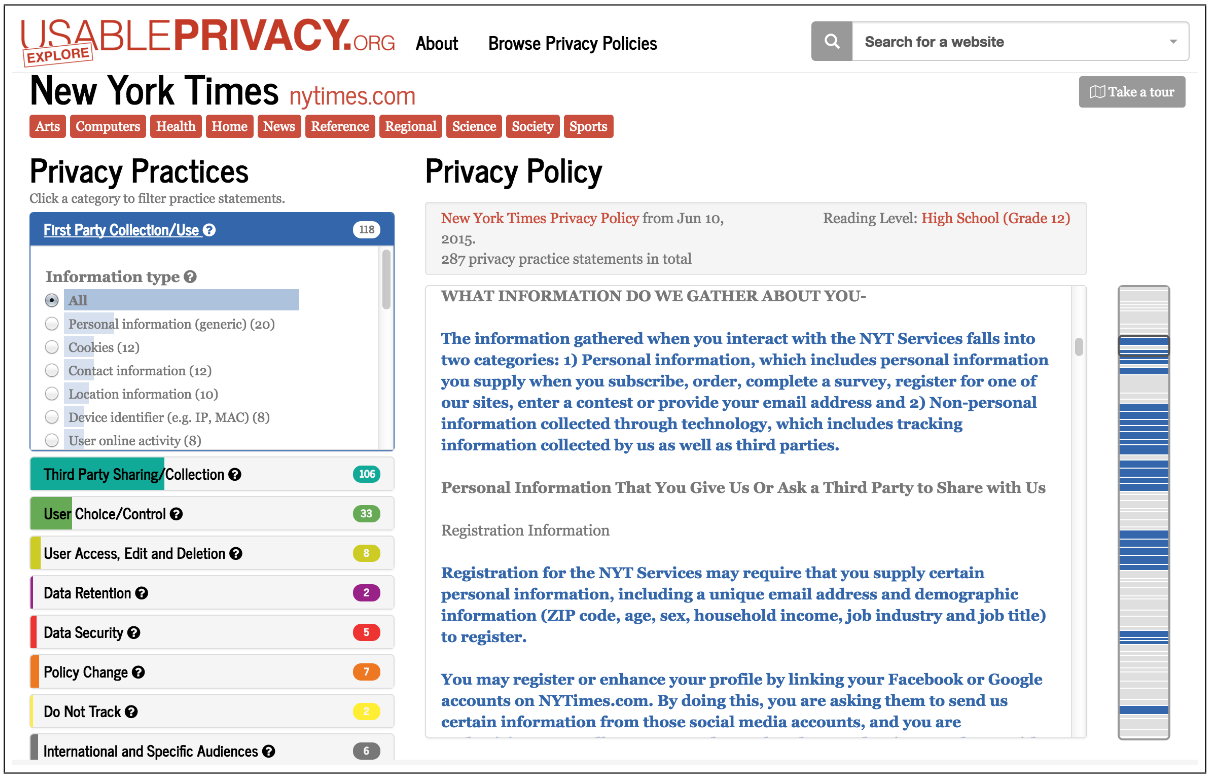

Fig. 11. A screenshot of the UPP Explore website that visualizes the First Party collection data practice of the New York Times' privacy policy.

[11] Duma, C., Herzog, A., and Shahmehri, N.: Privacy in the semantic web: What policy languages have to offer. POLICY '07: IEEE Int. Workshop Policies for Dist. Sys. Netw., 109-118 (2007)

[12] Ashley, P., Hada, S., Karjoth, G., Powers, C., and Schunter, M.: Enterprise privacy authorization language (EPAL 1.2). Submission to W3C, 156. (2003)

[13] Fillmore, C. J.: Frame semantics and the nature of language. Annals of the New York Academy of Sciences 280, no. 1, 20-32. (1976)

[14] Federal Trade Commission: Privacy Online: Fair Information Practices in the Electronic Marketplace. Report to Congress (2000)

[15] Cuenca Grau, B.: Privacy in ontology-based information systems: A pending matter. Semantic Web 1, 2, 137-141 (2010)

[16] Gandon, F.L., and Sadeh, N.M.: Semantic web technologies to reconcile privacy and context awareness. Web Semantics: Science, Services and Agents on the World Wide Web. 241-260 (2004)

[17] Google: Google Trends. Accessed: March 15, 2016.

[18] Heurix, J., Zimmermann, P., Neubauer, T.: A taxonomy for privacy enhancing technologies. Computers & Security 53, 1-17 (2015)

[19] HIPAA Privacy Rule. 45 CFR Part 160 (2002)

[20] Hobbs, J. R. and Pan, F.: An Ontology of Time for the Semantic Web. ACM Transactions on Asian Language Processing (TALIP) Vol. 3, No. 1, pp. 66-85. (2004)

[21] Hoke, C., Cranor, L.F., Leon, P.G., and Au, A.: Are They Worth Reading? An In-Depth Analysis of Online TrackersâĂŹ Privacy Policies. I/S: J Law Pol. Info. Soc. (2015)

[22] Jensen, C. and Potts, C.: Privacy policies as decision-making tools: an evaluation of online privacy notices. Proc. SIGCHI conference on Human factors in computing systems (CHI), 471–578 (2004)

[23] Jutla, D.N, Bodorik, P., and Zhang, Y.: PeCAN: An architecture for usersâĂŹ privacy-aware electronic commerce contexts on the semantic web. Information Systems 31, 4-5 (2006)

[24] Kagal, L., Berners-Lee, T., Connolly, D., and Weitzner, D.: Using semantic web technologies for policy management on the web. Proceeding of the National Conference on Artificial Intelligence, Vol. 21. No. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press (1999)

[25] Kaufmann, E., and Abraham B: Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases. Web Semantics, Science, Services and Agents on the World Wide Web 8.4, 377-393 (2010)

[26] Kost, M., Freytag, Christoph, Kargl, F., Kung, A.: Privacy verification using ontologies. ARES '11: Int. Conf. Availability, Reliability and Security, IEEE (2011)

[27] Lorch, M., Proctor, S., Lepro, R., Kafura, D., and Shah, S.: First experiences using XACML for access control in distributed systems. In ACM workshop on XML security (2003)

[28] McDonald, A. M., and Cranor, L.F.: The Cost of reading privacy policies. I/S J Law & Policy Info. Soc., 4(3) (2008)

[29] Negroponte, N.: Being Digital Random House Inc., New York, NY, USA (1995)

[30] Niles, I. and Pease, A.: Origins of the IEEE standard upper ontology. In Working notes of the IJCAI-2001 workshop on the

IEEE standard upper ontology, pp. 37-42. (2001).
[31] Online Privacy Protection Act of 2003. California Business and Professional Code, 22575–22579 (2004)
[32] Parsia, B., and Evren, S.: Pellet: An OWL DL reasoner. Int. Semantic Web Conf., Poster (2004)
[33] President's Concil of Advisors on Science and Technology: Big Data and Privacy: a Technological Perspective. Executive Office of the President, USA (2014).
[34] Reidenberg, J.R., Bhatia, J., Breaux, T.D., and Norton, T.B.: Automated Comparisons of Ambiguity in Privacy Policies and the Impact of Regulation. J Legal Studies 47 (forthcoming).
[35] Reidenberg, J.R., Breaux, T.D., Cranor, L.F., French, B., Grannis, A., Graves, J.T., Liu, F., McDonald, A.M., Norton, T.B. Ramanath, R., Russell, N.C., Sadeh, N., and Schaub, F.: Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding. Berkeley Technology Law Journal 30, 1 (2015)
[36] Reidenberg, J.R., Russell, N.C., Callen, A.J., Qasir, S., and Norton, T.B.: Privacy harms and the effectiveness of the notice and choice framework. I/S: J. Law & Policy Info. Soc. 11, 2 (2015)
[37] Renars, L., Cerans, K., and Sprogis, A: Visualizing and Editing Ontology Fragments with OWLGrEd. I-SEMANTICS (Posters & Demos) (2012)
[38] Sackmann, S., and Kahmer, M.: ExPDT: A policy-based approach for automating compliance. Wirtschaftsinformatik 50.5, 366 (2008)
[39] Sadeh, N., Gandon, F., and Oh Buyng, K.: Ambient Intelligence: The MyCampus Experience. In Ambient Intelligence and Pervasive Computing. Eds. T. Vasilakos and W. Pedrycz, ArTech House (2006)
[40] Schaub, F., Balebako, R., Durity, A.L., and Cranor, L.F.: A Design Space for Effective Privacy Notices. In Symposium on Usable Privacy and Security (2015)
[41] Shvaiko, P., Oltramari, A., Cuel, R., and Pozza, D: Generating innovation with semantically enabled TasLab portal. The Semantic Web: Research and Applications. Springer Berlin Heidelberg, 348–363 (2010)
[42] Toninelli, A., Montanari, R.,Kagal,L., and Lassila, O.: Proteus: A Semantic Context-Aware Adaptive Policy Model. In POLICY '07: Int. Workshop Policies Distr. Sys. Netw, IEEE (2007)
[43] Tonti, G., Bradshaw, J.M, Jeffers, R., Montanari, R., Suri, N. and Uszok A.: Semantic Web Languages for Policy Representation and Reasoning: A Comparison of KAoS, Rei, and Ponder. Int. Semantic Web Conf. (2003)
[44] Uszok, A., et al.: KAoS policy and domain services: Toward a description-logic approach to policy representation, deconfliction, and enforcement. In POLICY '03: Int. Workshop Policies Distr. Sys. Netw, IEEE (2003)
[45] Wilson, S., Schaub, F., Ramanath, R., Sadeh, N., Liu, F., Smith, N.A., and Liu, F.: Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work? In WWW '16: Int. World Wide Web Conf. (2016)
[46] Wilson, S., Schaub, F., Dara, A., Liu, F., Cherivirala, S., Leon, P., Andersen, M., Zimmeck, S., Sathyendra, K., Russell, N. C., Norton, T. B., Hovy, E., Reidenberg, J., and Sadeh, N. The creation and analysis of a website privacy policy corpus. In Proc. ACL 2016, Berlin, Germany, August 2016.
[47] Zimmeck, S., Wang, Z., Zou, L., Iyengar, R., Liu, B., Schaub, F., Wilson, S., Sadeh, N., Bellovin, S. M., and Reidenberg, J. Automated Analysis of Privacy Requirements for Mobile Apps, In Proc. NDSS 2017, San Diego, CA, February 2017.
[48] Liu, F., Wilson, S., Schaub, F., and Sadeh, N. Analyzing Vocabulary Intersections of Expert Annotations and Topic Models for Data Practices in Privacy Policies. In Proc. AAAI FSS: Privacy and Language Technologies, November 2016.
[49] Sathyendra, K., Schaub, F., Wilson, S., and Sadeh, N. Automatic Extraction of Opt-Out Choices from Privacy Policies. In Proc. AAAI FSS: Privacy and Language Technologies, November 2016.
[50] Breaux, T. D., Hibshi, H., and Rao, A. A formal language for specifying and analyzing data flow specifications for conflicting privacy requirements. Requirements Engineering, pages 1-27, 2013.
[51] Sadeh, N., Acquisti, A., Breaux, T., Cranor, L., McDonald, A., Reidenberg, J., Smith, N., Liu, F., Russell, N. C., Schaub, F., Wilson, S., Graves, J., Leon, P., Ramanath, R., and Rao, A. Towards usable privacy policies: Semi-automatically extracting data practices from websites' privacy policies. In Proc. SOUPS 2014.
[52] Acquisti, A. Nudging privacy: The behavioral economics of personal information. IEEE Security & Privacy, 7(6):82-85, 2009.
[53] Kelley, P. G., Bresee, J., Cranor, L. F., and Reeder, R. W. A 'Nutrition label' for privacy. In Proc. SOUPS '09. ACM, 2009.
[54] Gharib, M. and Giorgini, P. and Mylopoulos, J., Ontologies for Privacy Requirements Engineering: A Systematic Literature Review. ArXiv preprint arXiv:1611.10097, 2016.