Semantic Approach to Automating Management of Big Data Privacy Policies

Karuna P Joshi, Aditi Gupta, Sudip Mittal, Claudia Pearce, Anupam Joshi, and Tim Finin
CSEE Department
University of Maryland, Baltimore County
Baltimore, USA
{karuna.joshi, adigupta, smittal1, cpearce, joshi, finin}@umbc.edu

***Abstract*—Ensuring privacy of Big Data managed on the cloud is critical to ensure consumer confidence. Cloud providers publish privacy policy documents outlining the steps they take to ensure data and consumer privacy. These documents are available as large text documents that require manual effort and time to track and manage. We have developed a semantically rich ontology to describe the privacy policy documents and built a database of several policy documents as instances of this ontology. We next extracted rules from these policy documents based on deontic logic which can be used to automate management of data privacy. In this paper we describe our ontology in detail along with the results of our analysis of privacy policies of prominent cloud services.**

Keywords-data privacy; Personally Identifiable Information; semantic web; cloud computing; policy automation

I. INTRODUCTION

With the promise of rapid provisioning, scalability and high computing capability, cloud platforms are being adopted as the default computing environment for Big Data analytics and processing. Cloud-based service providers are collecting large amounts of data about their consumers including Personally Identifiable Information (PII) like contact addresses, credit card details, bank account details etc. Additionally, with the proliferation of loyalty reward cards, retailers are also tracking the buying patterns of their consumers and often storing this large dataset on public cloud-based systems.

While data privacy is critical for all Big Data applications, ensuring privacy of data stored on cloud platforms is more challenging since large chunks of data pass through the global Internet and the high availability (multiple data replication) and device independence features of the cloud make it more vulnerable to data breaches. Hence the user community of Big Data on the cloud are very concerned about data privacy controls. Some of the key privacy concerns of Big Data consumers identified by Brill [2] are summarized below

- "De-Identified" Information Can Be "Re-Identified": Data collectors claim that the aggregated information has been "de-identified", however, it is possible to re-associate "anonymous" data with specific individuals, especially since so much information is linked with smartphones.
- Possible Deduction of PII: Non-personal data could be used to make predictions of a sensitive nature, like sexual orientation, financial status, etc.
- Risk of Data Breach Is Increased: The higher the concentration of data, the more appealing a target it makes for hackers, and the greater impact as a result of the breach.
- "Creepy" Factor: Consumers are often unnerved when they feel that companies know more about them than they are willing to volunteer. For instance, the shopping history tracked by large retailers can be used to predict consumer lifestyle choices.
- Big Brother or Big Data: Municipalities are using Big Data for predictive policing and tracking potential terrorist activities. Concerns have been raised that such uses could become a slippery slope to using Big Data in a manner that infringes on individual rights, or could be used to deny consumers important benefits (such as housing or employment) in lieu of credit reports.

To address these concerns, government agencies and regulatory bodies around the world have developed policies and guidelines to secure cloud data. Cloud providers also provide consumers with privacy policy documents that describe their privacy controls in detail. These documents are an essential component of their service contract. However, these documents are often text based and require manual effort to parse and manage. A critical step in automating privacy management is to make the privacy documents machine processable so that monitoring tools can interpret the policy rules and metrics defined in them. We have developed a semantically rich approach to automate the management of privacy policy documents, using Semantic Web technologies, Natural Language Processing (NLP) and text mining techniques. We describe our privacy ontology along with the results we obtained by analyzing privacy documents of 10 prominent cloud service providers. Our work can be used by Big Data consumers for writing data privacy policies using formal policy languages and building automated systems for compliance validation.

In this paper, we initially discuss the background and related work in this area. In section III, we describe our approach towards automating data privacy documents and describe the ontology we have developed for the same using OWL[9]. In section IV and V we describe the text mining and NLP approaches we took to extract and populate privacy policy documents of various cloud-based service providers as instances of our ontology and the results of our analysis of privacy policy documents. We end with conclusions and future work.

II. Related Work

A. Privacy Features

While identifying the critical privacy controls that should be specified by the privacy policy documents, we reviewed various standards and guidelines proposed for data privacy policy by organizations like National Institute of Standards and Technology (NIST) [1][3][4], European Union data protection standard [43][44], privacyalliance.org [5], Federal Trade Commission [6], and the United States Small Business Administration [7]. All of them have identified the following topics as key points to be covered:

1. Explain how a business collects and uses personal information: This includes cookie policy, contact information and how the customer data is shared and with whom.
2. Age restrictions and collecting data from children – If a business targets children under the age of 13, they'll need to comply with the Children's Online Privacy Protection Act (COPPA) in the US.
3. Displaying the privacy policy and notice of data collection – A business has to make sure that new customers or users have easy access to their policy by prominently displaying links to it. It must also mention the fact that user's data might be collected. All visitors and users of a website must be notified of the Privacy Policy before any personal or identifiable data is collected.
4. Purpose and consent of data collection: businesses should list all purposes for which they will use collected data or information in the Privacy Policy.
5. Security of collected data: All businesses must commit to the security and safeguarding of user data collected or submitted on their websites. A Privacy Policy must address any potential data-security concerns users or visitors of a website may have. The Policy must also outline all steps used to secure identifiable data or information and ensure that it is safeguarded at all times.
6. User access to data: businesses must always make personally identifiable information accessible to users of their websites. Users must be able to edit, change or delete information from the website at any time in a simple and straightforward manner.

These guidelines are vital in understanding a privacy policy and provide a shell that can be used to extract knowledge from an unstructured privacy policy document. Another important pattern that emerges from various publicly-available privacy policies is the fact that various providers in our study, as well as others not included in our study, follow a 'topic – description' model of describing their privacy policies where a topic heading is listed above a few sets of paragraphs that provide details about that topic.

B. Semantic Web

In a virtualized service-oriented environment, consumers and providers need to be able to exchange information, queries, and requests with some assurance that they share a common meaning. This is critical not only for the data but also for the policies followed by service consumers or providers. The handling of heterogeneous policies is usually not present in a closed and/or centralized environment, but is an issue in the open cloud. The interoperability requirement is not just for the data itself, but even for describing services, their service level agreements, quality related measures, and their policies for sharing data.

One possible approach to this issue is to employ Semantic Web techniques for modeling and reasoning about services related information. We have used this approach for automating cloud privacy policy documents. The Semantic Web deals primarily with data instead of documents. It enables data to be annotated with machine understandable meta-data, allowing the automation of their retrieval and their usage in correct contexts. Semantic Web technologies include languages such as Resource Description Framework (RDF) [8] and Web Ontology Language (OWL) [9] for defining ontologies and describing meta-data using these ontologies as well as tools for reasoning over these descriptions. These technologies can be used to provide common semantics of privacy information and policies enabling all agents who understand basic Semantic Web technologies to communicate and use each other's data and Services effectively.

In one of our prior works, we described a new integrated methodology for the lifecycle of IT services delivered on the cloud, and demonstrate how it can be used to represent and reason about services and service requirements and so automate service acquisition and consumption from the cloud. We have divided the IT service lifecycle into five phases of requirements, discovery, negotiation, composition, and consumption. We detail each phase and describe the ontologies that we have developed to represent the concepts and relationships for each phase. We have described the five phases in detail along with the associated metrics in [39].

C. Text Extraction

Researchers have applied NLP techniques to extract information from text documents. In Rusu et. al. [10] the authors suggest an approach to extract subject-predicate-object triplets. They generate Parse Trees from English sentences and extract triplets from the parse trees. Etzioni et. al. [11] developed the KNOWITALL system to automate the process of extracting large collections of facts from the Web in an unsupervised, domain-independent, and scalable manner. Etzioni et. al used Pattern Learning to address this challenge. Various textual information extraction and retrieval systems have been proposed in [12][13][14][15].

Another important NLP technique used for information extraction from unstructured text is 'Noun Phrase Extraction'. Rusu et. al. in [10] show how to create triplets by considering 'Noun Phrases' obtained by using various part-of-speech taggers. Barker et. al. [16] extract key-phrases from documents and show that noun phrase-based system performs roughly as well as a state of the art, corpus-trained key-phrase extractor. Similar techniques have also been suggested in [11].

Use of automated techniques for extracting permissions and obligations from legal documents, such as text mining and semantic techniques have been explored by researchers in the past [17][18][19]. Kagal et al. [20][21] have proposed an ontology based policy framework to model conversation specifications and policies using obligations and permissions.

III. TECHNICAL APPROACH

In this paper, we utilize Semantic Web, NLP and text mining techniques to semi-automate the process of knowledge extraction from privacy documents. We identified the key information and controls to be included in a privacy policy and defined a formal and extensive ontology to represent these controls.

We followed a three phase approach to build a semantic framework for analyzing privacy documents. We used text mining approach to extract and populate the semantic knowledge graph. For our analysis we created a corpus of publicly available privacy policies of companies like Amazon AWS [22], Facebook [22], Google [24], HP [25], Oracle [26], PayPal [27], Salesforce [28], Snapchat [29], Twitter [30], WhatsApp [31]. We stored the knowledge extracted from these documents as RDF tuples. Following are the three phases:

1. **Ontology Development**: We defined a detailed ontology for cloud privacy documents using OWL language. This is described in detail in section IV.
2. **Extracting terms and definitions**: We have previously developed a prototype system to automatically extract key definitions and measures from the legal cloud documents [37][38]. We used this system to extract key terms and topics from the privacy documents of the vendors listed above. Section V describes this approach and the results we obtained.
3. **Analyzing permissions and obligations rules**: We proposed and evaluated techniques to analyze different privacy documents based on Deontic Logic formalizations. We extracted rules with deontic modalities and tagged them as obligations and permissions. This is described in detail in section VI.

In order to extract various terms, definitions, permissions and obligations, we convert sentences into parse trees using technologies like the CMU Link Parser [34] and then use various rules to mine key information from these parse trees. For more details, see Section V.

IV. ONTOLOGY FOR PRIVACY POLICY DOCUMENTS

We have developed a detailed ontology using semantic web language OWL to define the range of information that should be included in the Privacy Policy documents.

On reviewing the privacy policies of leading cloud service providers listed in section III, we observed that they primarily describe the user data they capture and use and/or share. We compared the various data privacy standards that will be best suited for Big Data applications hosted on the cloud and determined that NIST Special Publication 800-144 [3], that provides guidelines on security and privacy in public cloud computing, and NIST SP 800-53 [4], that listed the privacy controls that are part of the federal cloud computing standards, are best suited for our ontology. These privacy controls are based on the Fair Information Practice Principles (FIPPs) 121 embodied in the Privacy Act of 1974, Section 208 of the E-Government Act of 2002, and Office of Management and Budget policies. As part of our ongoing work we are working on expanding the ontology to cover other privacy controls. This ontology is available in the public domain and can be accessed at [46]

A. Privacy Controls included in Ontology

In this paper we concentrate on the three families of privacy control, identified in NIST SP 800-53 [4] and listed below, that are relevant to all organizations. Many state laws require web service providers to display their privacy policies and procedures [45].

1) Authority and Purpose

- **Authority to Collect**: The service provider should determine and document the legal authority that permits the collection, use, maintenance, and sharing of personally identifiable information (PII), if required by regulatory and compliance bodies.
- **Purpose Specification Control**: The organization describes the purpose(s) for which personally identifiable information (PII) is collected, used, maintained, and shared in its privacy notices.

2) Transparency

This family ensures that organizations provide public notice of their information practices and the privacy impact of their programs and activities.

- **Privacy Notice**: The organization
  a. provides notice to the public and to individuals regarding (i) its activities that impact privacy, including its collection, use, sharing, safeguarding, maintenance, and disposal of PII; (ii) authority for collecting PII; (iii) the choices, if any, individuals may have regarding how the organization uses PII and the consequences of exercising or not exercising those choices; and (iv) the ability to access and have PII amended or corrected, if necessary;
  b. Describes: (i) the PII the organization collects and the purpose(s) for which it collects that information; (ii) how the organization uses PII internally; (iii) whether the organization shares PII with external entities, the categories of those entities, and the purposes for such sharing; (iv) whether individuals have the ability to consent to specific uses or sharing of PII and how to exercise any such consent; (v) how individuals may obtain access to PII; and (vi) how the PII will be protected; and
  c. Revises its public notices to reflect changes in practice or policy that affect PII or changes in its

activities that impact privacy, before or as soon as practicable after the change.

- **Dissemination of Privacy Program Information**: The organization ensures that the public has access to information about its privacy activities and is able to communicate with its Senior Agency Official for Privacy (SAOP)/Chief Privacy Officer (CPO); and ensures that its privacy practices are publicly available through organizational websites or otherwise.

3) Use Limitation

This family ensures that organizations only use PII either as specified in their public notices, in a manner compatible with those specified purposes, or as otherwise permitted by law.

- **Internal Use:** The organization uses PII internally only for the authorized purpose(s) identified in the Privacy Act and/or in public notices.
- **Information Sharing with Third Parties**: The organization:
  a. Shares PII externally only for the authorized purposes identified in the Privacy Act and/or described in its notice(s) or for a purpose that is compatible with those purposes;
  b. Where appropriate, enters into Memoranda of Understanding, Memoranda of Agreement, Letters of Intent, Computer Matching Agreements, or similar agreements, with third parties that specifically describe the PII covered and specifically enumerate the purposes for which the PII may be used;
  c. Monitors, audits, and trains its staff on the authorized sharing of PII with third parties and on the consequences of unauthorized use or sharing of PII;
  d. Evaluates any proposed new instances of sharing PII with third parties to assess whether the sharing is authorized and whether additional or new public notice is required.

B. *Privacy Policy Ontology*

The main classes of the ontology are illustrated in figure 1. Referring to the NIST guidelines on cloud privacy [3] and PII information [1], we have identified the key components of a privacy notice that are defined as object properties in the main Privacy Policy class. The numbers in the brackets indicate the relationship with the class of the functional property. So each privacy policy should have one instance describing the collection purpose and data protection controls; privacy policy should have at least one instance of consumer consent and Access to own PII, and so on. The main sub-classes are -

1) Collection Purpose

This class captures the purpose and scope of data collection and the limited use that the data will be subjected to. It also contains information of the actions that will be taken to transform the data, which can include combining it with other datasets or aggregating/summing the data. The policy document should also specify the duration the data will be managed by the data collector and the deletion and archival actions that will be taken after that duration ends.

2) PII Data Collected

This class identifies key attributes that comprise personal identifiable information. These include personal details like

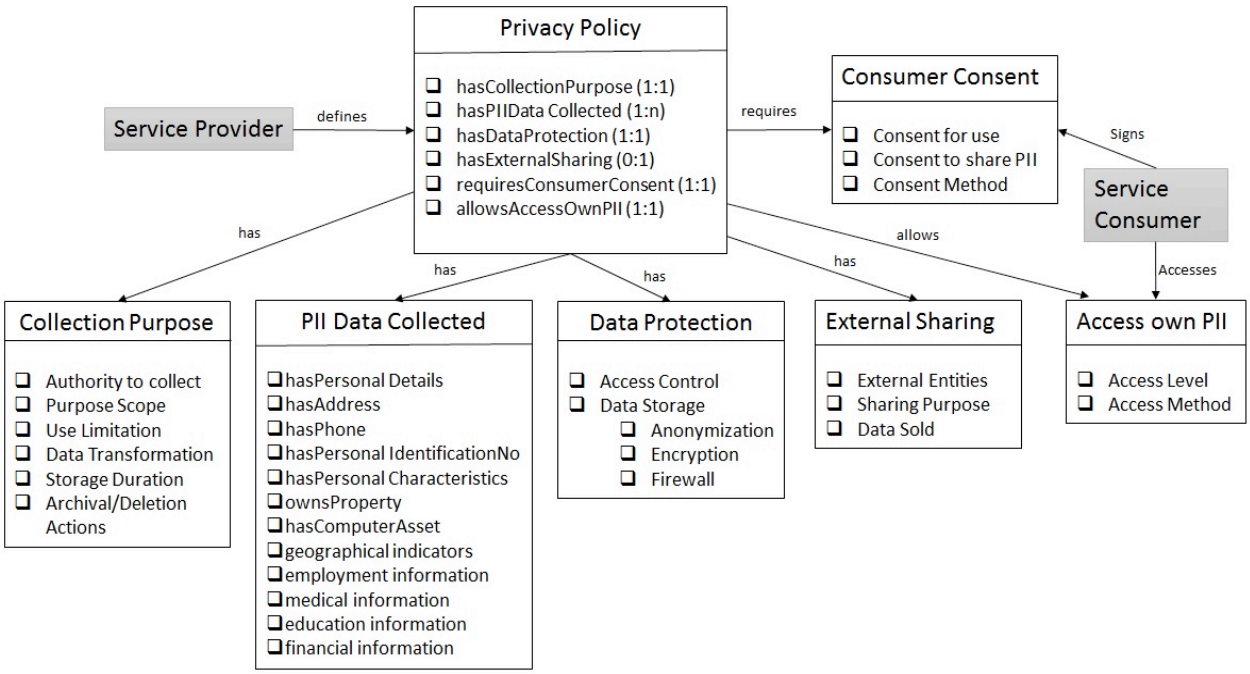

Figure 1: Top Level Ontology describing components of data privacy policy

names, contact information, like address, phone numbers, identity numbers and identity characteristics. These are illustrated in detail in figure 2. Other PII data includes employment, medical, financial and education details of a person. To identify the key properties of these classes, we referenced the NIST special publication 800-122. [1]

Each privacy policy instance may have one or more instances of PII data associated with it, but the number of instances should be small. We allow multiple instances of PII to accommodate data versioning and allow keeping old values of PII even when new value is added. For instance, a consumer may change their primary address associated with an e-commerce site, the site vendor could retain the previous address of the consumer in a separate instance for internal analysis of consumer behavior. Alternatively, the provider might want to change the PII dataset collected by their service, but retain the same collection purpose and data protection policies and so would have multiple instances of just the PII Data Collected class.

As part of our ongoing work, we are linking this ontology with other existing ontologies in the public domain. For instance, the geographical indicators will be linked with the W3C Geospatial Ontologies [41], financial information will reference EDM Council's FIBO [42] financial ontology; the medical information class will reference existing medical ontologies available at openclinical.org/ontologies.html, etc. For such classes a single block is displayed in figure 2.

3) Data Protection

This class includes properties pertaining to data access control and data storage controls that should be in place. In our previous work, we have developed OWL ontologies for Role based access control [36] and attribute based access control [35] which we plan to integrate with this privacy policy ontology. As part of our planned work, we will also incorporate other publically available data protection ontologies.

4) External Sharing

This class includes the details of external entities with whom the data will be shared. It includes the purpose of sharing this data and information about whether the data will be sold to these external entities. This is the only class storing non mandatory information. If the data collected will not be shared externally, then there will be zero instances of this class which we have indicated as (0:1) in the Privacy policy class.

5) Consumer Consent

The consumer's consent should be obtained whenever PII data is captured by the provider. The consent to share the data should be explicitly mentioned. The consent method – signature, agreement etc. should be specified.

6) Access own PII

The consumer should be able to access their PII that is maintained by the provider. The access method should be

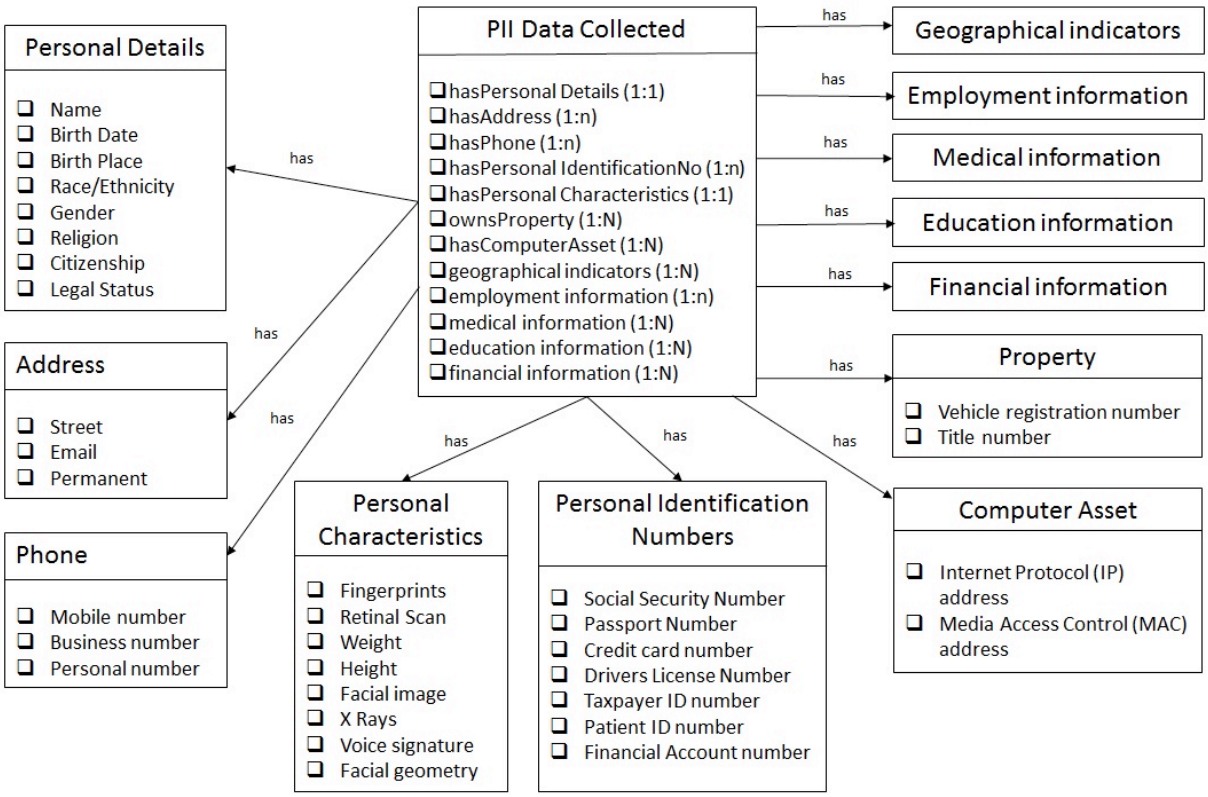

Figure 2: Details of the PII Data Collected class describing PII data items

clearly specified in the privacy document.

V. KNOWLEDGE EXTRACTION

In this section we describe our work on extracting knowledge or important details from privacy documents. For our analysis we created a corpus of publicly available privacy policies of 10 companies, viz. Amazon AWS [22], Facebook [22], Google [24], HP [25], Oracle [26], PayPal [27], Salesforce [28], Snapchat [29], Twitter [30] and WhatsApp [31]. We store the knowledge extracted from these documents in RDF format.

Extracting knowledge from unstructured privacy policies and representing it using RDF is advantageous as this knowledge can be used by various intelligent systems to automatically enforce system rules. An intelligent system can potentially discover the best possible deal a service user can procure based on user's privacy policy restrictions.

Privacy policy documents currently do not have a standard format and are defined by the service provider for each service. In our analysis of the 10 providers we have observed that while some providers provide a separate privacy policy document for each of their services, some have a single privacy policy notice for all their services, (for instance, Amazon AWS privacy policy document cross references the privacy notice listed on Amazon.com). Still other vendors (like WhatsApp) include it as part of their overall service contract. Due to this discrepancy in the format of private policy notices, analysis of these documents required us to perform a lot of manual pre-processing to identify specific sections in the document that should be included in the analysis so as to not skew the final result.

A. *Extraction Process*

In order to extract data from various privacy policy documents, we created a prototype system to automatically extract various details like definitions, PII, data sharing details, security details various restrictions, etc. Figure 3 illustrates the architecture diagram of this prototype. We begin by retrieving publicly-available privacy policy documents that are posted by various businesses on their websites. We pass these documents to our 2 modules, key-terms extractor and a topic-description extractor. We save the knowledge that we extract from these documents as RDF statements.

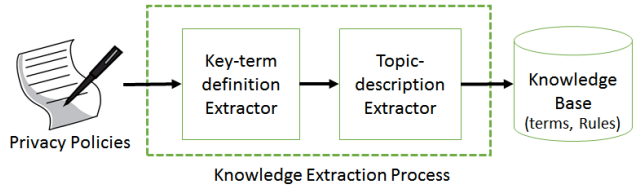

Figure 3: Architecture Diagram of Knowledge Extraction Process

1) *Key-terms extractor:*

Privacy policy documents created by various companies discuss various company/application specific key-terms. These terms are vital from a communication point of view and are highly relevant as they help a user understand a privacy document. For example, in Google's privacy policy 'affiliate' is a key term which is defined as 'An affiliate is an entity that belongs to the Google group of companies'. Similarly, 'Browser web storage' is defined as 'Browser web storage enables websites to store data in a browser on a device. When used in "local storage" mode, it enables data to be stored across sessions (for example, so that the data are retrievable even after the browser has been closed and reopened). One technology that facilitates web storage is HTML 5'.

In order to extract such key-terms and their definitions we use pattern learning. This technique involves learning a few extraction patterns and then using them to extract key-terms and their definitions. We divide the privacy policy documents into different sentences and then pass them through the CMU Link Parser [34]. The link parser generates a parse tree which is compared to various patterns. Each sentence that fits this pattern consists of a noun phrase, a connector and a verb phase. A typical pattern is: "X is a Y" where X is the noun phase, Y is the verb phase and 'is a' the connector. So in the example of the key-term 'affiliate' mentioned above we are able to extract the said key-term and its definition (illustrated in figure 4).

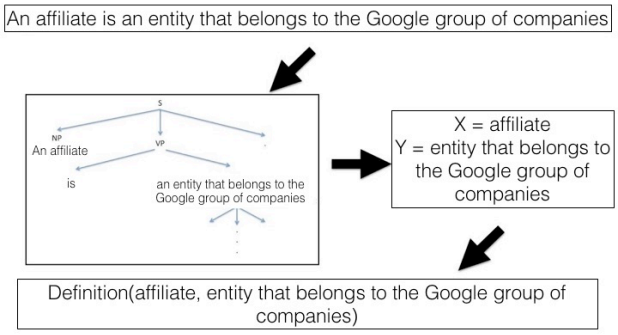

Figure 4: Noun Phrase Extraction when run on a definition in Google's privacy policy.

2) *Topic-description extractor*

In the second module we try to extract various fine details mentioned in privacy policy documents like, personal information collected, third party sharing, security details, age restrictions, condition of use, etc. So as to extract these details from unstructured text we use a topic-description 'noun-phrase extraction' approach. In the topic – description model various businesses mention a heading to a few sets of paragraphs and in these paragraphs provide more details about various important topics as per the guidelines of writing a privacy policy (see Section II A).

In this module we parse the document for certain topics of importance like, third party, security, etc. Once we find the location of these topics of importance, we isolate the paragraph where the said details are available. We then pass individual sentences present in these paragraphs through the CMU link parser to generate a parse tree. Using this parse tree, we look at various noun phrases present and

match them to our ontology to create triples. For example, in the privacy policy of Amazon AWS, '*Amazon.com, Inc. and its controlled U.S. subsidiaries, including Amazon Web Services, Inc., are participants in the Safe Harbor program developed by the U.S. Department of Commerce and (1) the European Union and (2) Switzerland, respectively.*' is a sentence discussing various details about the safe harbor program. We are able to create RDF triples about Amazon's involvement in the program using noun-phrase extraction.

B. Extraction Results

In this paper we analyzed privacy policies of 10 vendors listed above and extracted various details from these documents. Table 1 lists some key-terms automatically extracted from Google's privacy policy.

Key Term	Definition
Affiliate	an entity that belongs to the Google group of companies.
application data cache	a data repository on a device.
Cookie	a small file containing a string of characters that is sent to your computer when you visit a website.
Device	a computer that can be used to access Google services.
HTTP Referrer	information transmitted to a destination webpage by a web browser, typically when you click a link to that webpage
Internet protocol (IP) address	Every device connected to the Internet is assigned a number.
Non-personally identifiable information	information that is recorded about users so that it no longer reflects or references an individually identifiable user.
Personal information	information which you provide to us which personally identifies you, such as your name, email address or billing information, or other data which can be reasonably linked to such information by Google, such as information we associate with your Google account
pixel tag	a type of technology placed on a website or within the body of an email for the purpose of tracking activity on websites, or when emails are opened or accessed, and is often used in combination with cookies.
Sensitive personal information	a particular category of personal information relating to confidential medical facts, racial or ethnic origins, political or religious beliefs or sexuality.
server logs	typically include your web request, Internet Protocol address, browser type, browser language, the date and time of your request and one or more cookies that may uniquely identify your browser.
unique device identifier	a string of characters that is incorporated into a device by its manufacturer and can be used to uniquely identify that device

Table 1: Some key-term definition extracted from Google's privacy policy.

Table 2 lists the number of statements extracted from different policies.

Privacy Policy	Statements Extracted
Amazon AWS	72
Facebook	261
Google	333
HP	310
Oracle	291
PayPal	142
Salesforce	175
Snapchat	92
Twitter	177
WhatsApp	204

Table 2: Number of statements extracted from different privacy policies.

VI. Rules Extraction Using Permission and Obligation

In the previous sections of this paper, we have defined and extracted the various components of the knowledge graph based on the privacy policies of different service providers. Now, we use text mining and NLP techniques to extract relevant information in the form of deontic rules of permissions and obligations. These rules define the rights, obligations and prohibitions for the key stakeholders such as the service provider, users and third party entities. Extraction of these rules is essential in building a reasoning module over our ontology for automated management of privacy policies. Privacy policies contain details such as what PII is collected and under what conditions the service provider can use and share the PII information of the user, as well as the user's obligation and agreement to the service provider's policy. Analyzing and comparing these rules across different service providers can give the users useful insights about the privacy policies and ensure that the privacy rules are in compliance of the user's privacy needs. In one of our earlier works, we have used a similar approach in extraction of rules and obligations from service level agreements of cloud service providers [40].

A. Theory of Modal / Deontic Logic

Modal logic is a broad term used to cover various other forms of logic such as temporal logic and deontic logic [32]. Deontic logic describes statements containing permissions and obligations, and temporal logic describes time based requirements. Deontic logic further consists of four types of modalities:

1. **Permissions** / Rights: Permissions are expressions or rules that describe the rights or authorizations for an entity.
2. **Obligations**: Obligations expressions are the mandatory actions that an entity must perform.
3. **Dispensations**: Dispensations that describe optional expressions and describe non-mandatory conditions.

4. **Prohibitions**: Prohibitions are the expressions that specify the actions which are prohibited.

B. *Extraction of modalities using NLP techniques*

In order to extract modal expressions from the privacy policies, we used the Stanford Parser [32] to obtain the part-of-speech (POS) tags for each of the statements in the documents. Next we formulated grammatical rules based on the POS tags to obtain rules in the form of permissions and obligations. In our paper, we refer to deontic modalities to include both deontic and temporal logic statements. A sample of grammatical rules is given below:

Permissions / Obligations:
<Actor> <deontic modal> <verb>
<Actor> <deontic modal> <adverb> <verb>

Prohibitions / Dispensations:
<Actor> <deontic modal> <negation> <verb>

Deontic modals used:
- Obligations and prohibitions: should, shall, must
- Permissions and dispensations: can, may, could
- Temporal modalities: will

C. *Results*

We used the 10 privacy policy documents from various service providers listed in the previous section for our analysis [1]. We present the analysis of the deontic and temporal modalities extracted using the grammatical rules defined.

1) *Extraction Results*

Using the grammatical expressions defined above we extracted the modal expressions and then categorized the statements as permissions, obligations, etc. based on the modal verbs present. The number of extracted statements varies across the documents depending upon the length and details provided in each of the documents. In total we extracted around 535 rules based on the deontic and modal logic formalizations. About 77% of extracted rules were permissions and 19% were temporal modalities, while the rest 4% were other categories. Figure 5 shows the distribution of the extracted rules across different service providers.

Some of the sample statements containing deontic expressions extracted from the privacy policies by our analysis include -

- "You can choose not to provide certain information, but then you might not be able to take advantage of many of our features." (Type: Permission, Actor: User)
- "We may also collect technical information to help us identify your device for fraud prevention and diagnostic purposes." (Type: Permission, Actor: Service Provider)

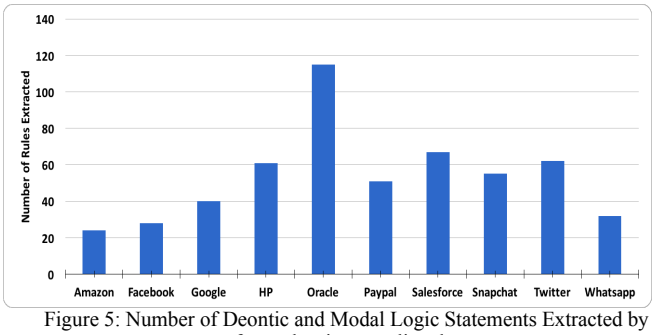

Figure 5: Number of Deontic and Modal Logic Statements Extracted by our system for each privacy policy document

2) *Actor analysis*

For each of the deontic statements extracted by our system we also automatically extracted the actor to which the permission and obligations applies. We use the noun / pronoun part of the part-of-speech tagging to assign actors for each of the deontic statements. We categorize the actors in three broad categories as shown in Figure 6. Majority (43%) of the modalities apply to the company or the service providers, while about 24% are for the users and the customers. The rest of the modalities belong to other actors such as third party services, partners and applications or remain uncategorized.

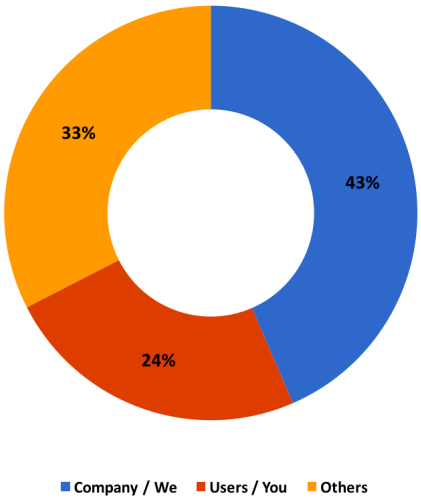

Figure 6: Distribution of actors for the modalities extracted. Majority of permissions, obligations and other modalities apply to the company or service providers.

3) *Verb-based analysis*

Analysis of the verbs in deontic modalities indicates the actions which are governed by the permissions and obligations. The tag cloud in figure 7 shows that most

[1] For Amazon, we consider the Amazon.com Privacy Notice, which is also referenced in the Amazon Web Services privacy policy. For WhatsApp we consider the Privacy Notice section of their Terms of Service document.

permissions and obligations were about actions such as: use, provide, collect and share the information and content. Hence, extraction and analyzing the deontic rules can give useful insights to users about the policies of the service provider regarding their information and content.

Figure 7: Analysis of Verbs in Deontic and Temporal Modalities Expressions

VII. Conclusion and Future Work

Currently privacy policy documents are managed as text files. As a result, extensive manual effort is required to monitor the privacy constraints on sharing and managing PII data. We have worked on significantly automating this process using semantic web technologies like OWL. In this paper we have described our ontology and text extraction techniques that we have developed to illustrate how the permissions and obligations can be automatically extracted from privacy policy document.

As part of our ongoing work, we are working with legal experts to validate and enhance our privacy document ontology. We are also working on linking this ontology with other existing ontologies in public domain. For instance, the geographical indicators will be linked with the W3C Geospatial Ontologies [41], financial information will reference EDM Council's FIBO [42] financial ontology; the medical information class will reference existing medical ontologies available at openclinical.org/ontologies.html, etc.

In the future we would like to add functionality which will allow users to compare and contrast privacy policies of different service providers. Some constraints in the privacy policy documents are listed in tables and hence are difficult to find using language based extractors; we would like to address this issue in the future. We are also building our dataset of privacy policy documents to be able to further refine our privacy policy ontology.

Acknowledgment

This research was partially supported by a DoD supplement to the NSF award #1439663: NSF I/UCRC Center for Hybrid Multicore Productivity Research (CHMPR).

References

[1] NIST Special Publication 800-122, Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) , http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf

[2] J. Brill, FTC Big Data Issues, 2012 https://www.ftc.gov/public-statements/2012/03/big-data-big-issues

[3] W. Jansen, T. Grance, NIST SP 800-144 Guidelines on Security and Privacy in Public Cloud Computing, http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-144.pdf

[4] NIST SP 800-53, http://disa.mil/services/dod-cloud-broker/~/media/files/disa/services/cloud-broker/nist-sp80053-securityandprivacycontrols.pdf

[5] Privacy Alliance, http://www.privacyalliance.org/resources/ppguidelines/

[6] Federal Trade Commission (FTC), https://www.ftc.gov/tips-advice/business-center/privacy-and-security

[7] U.S.S.B.A. (United States Small Business Administration) https://www.sba.gov/blogs/7-considerations-crafting-online-privacy-policy

[8] O. Lassila, R. Swick and others, Resource Description Framework (RDF) Model and Syntax Specification, WWW Consortium, 1999.

[9] D. McGuinness, F. Van Harmelen, et al., OWL web ontology language overview, W3C recommendation, World Wide Web Consortium, 2004.

[10] D. Rusu, L. Dali, B. Fortuna, M. Grobelnik, and D. Mladenic, "Triplet extraction from sentences," in Proceedings of the 10th International Multiconference" Information Society-IS, 2007, pp. 8–12.

[11] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Unsupervised namedentity extraction from the web: An experimental study," Artificial intelligence, vol. 165, no. 1, pp. 91–134, 2005.

[12] F. Ciravegna, "2, an adaptive algorithm for information extraction from web-related texts," in In Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining. Citeseer, 2001.

[13] S. Soderland, "Learning to extract text-based information from the world wide web." in KDD, vol. 97, 1997, pp. 251–254.

[14] P. Cimiano, S. Staab, and J. Tane, "Automatic acquisition of taxonomies from text: Fca meets nlp," in Proceedings of the International Workshop & Tutorial on Adaptive Text Extraction and Mining held in conjunction with the 14th European Conference on Machine Learning and the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2003.

[15] J. Cowie and W. Lehnert, "Information extraction," Communications of the ACM, vol. 39, no. 1, pp. 80–91, 1996.

[16] K. Barker and N. Cornacchia, "Using noun phrase heads to extract document keyphrases," in Advances in Artificial Intelligence. Springer, 2000, pp. 40–52.

[17] T. D. Breaux and A. I. Anton, "Analyzing goal semantics for rights, permissions, and obligations," in RE'05: Proceedings of the 13th IEEE International Requirements Engineering Conference (RE'05), IEEE Computer Society, August 2005, pp. 177–186.

[18] T. D. Breaux, M. W. Vail, and A. I. Anton, "Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations," in RE'06: Proceedings of the 14th IEEE International Requirements Engineering Conference (RE'06), IEEE Society Press, September 2006.

[19] N. Kiyavitskaya, N. Zeni, T. D. Breaux, A. I. Anton, J. Cordy, L. Mich, and J. Mylopoulos, "Automating the extraction of rights and obligations for regulatory compliance," in ER'08: Proceedings of the 27th International Conference on Conceptual Modeling (ER'08), Springer-Verlag, October 2008.

[20] L. Kagal and T. Finin, Agent Communication: International Workshop on Agent Communication, AC 2004, New York, NY, USA, July 19, 2004, Revised Selected and Invited Papers. Springer Berlin Heidelberg, 2005, ch. Modeling Communicative Behavior Using Permissions and Obligations.

[21] Kagal, L. & Finin, T., “Modeling conversation policies using permissions and obligations,” Auton Agent Multi-Agent Syst (2007) 14: 187. doi:10.1007/s10458-006-0013-z

[22] Amazon Privacy Notice: https://www.amazon.com/gp/help/customer/display.html?nodeId=468496

[23] Facebook Privacy Policy: https://www.facebook.com/policy.php

[24] Google Privacy Policy: https://www.google.com/policies/privacy/

[25] HP Privacy Policy: http://www8.hp.com/us/en/privacy/privacy.html

[26] Oracle Privacy Policy: https://www.oracle.com/legal/privacy/privacy-policy.html

[27] PayPal Privacy Policy, https://www.paypal.com/webapps/mpp/ua/privacy-full

[28] Salesforce Privacy Policy, http://www.salesforce.com/company/privacy/

[29] Snapchat Privacy Policy: https://www.snapchat.com/privacy

[30] Twitter Privacy Policy: https://twitter.com/privacy?lang=en

[31] WhatsApp Privacy Policy: https://www.whatsapp.com/legal/

[32] Modal Logic: http://plato.stanford.edu/entries/logic-modal/

[33] “The stanford parser: A statistical parser.” [Online]. Available: http://nlp.stanford.edu/software/lex-parser.shtml

[34] “Link grammar.” [Online]. Available: http://www.link.cs.cmu.edu/link/

[35] Nitin Kumar Sharma and Anupam Joshi, Representing Attribute Based Access Control Policies in OWL , ICSC 2016, DOI: http://doi.ieeecomputersociety.org/10.1109/ICSC.2016.16

[36] Tim Finin et al., "ROWLBAC - Representing Role Based Access Control in OWL", InProceedings, Proceedings of the 13th Symposium on Access control Models and Technologies, June 2008

[37] Sudip Mittal, Karuna Joshi, Claudia Pearce, and Anupam Joshi, Automatic Extraction of Metrics from SLAs for Cloud Service Management, in proceedings of IEEE International Conference on Cloud Engineering, April 2016 (IC2E 2016)

[38] Sudip Mittal, Karuna Joshi, Claudia Pearce, and Anupam Joshi, "Parallelizing Natural Language Techniques for Knowledge Extraction from Cloud Service Level Agreements.", poster at IEEE International conference on BigData, October, 2015

[39] Karuna P. Joshi, Yelena Yesha, Tim Finin, "Automating Cloud Services Lifecycle through Semantic Technologies," IEEE Transactions on Services Computing, vol.7, no.1, pp.109-122, Jan.-March 2014; doi: 10.1109/TSC.2012.41

[40] Aditi Gupta, Sudip Mittal, Karuna Pande Joshi, Claudia Pearce, and Anupam Joshi "Streamlining Management of Multiple Cloud Services", InProceedings, IEEE International Conference on Cloud Computing, June 2016.

[41] Joshua Lieberman, Raj Singh, Chris Goad, W3C Geospatial Ontologies, https://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/, retrieved 8/7/16

[42] Financial Industry Business Ontology (FIBO) , EDM Council, http://www.edmcouncil.org/financialbusiness, retrieved 8/7/16

[43] European Commission, Protection of Personal Data, http://ec.europa.eu/justice/data-protection/

[44] Regulation 2016/679 of the European parliament, http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC

[45] State Laws related to Internet Privacy, http://www.ncsl.org/research/telecommunications-and-information-technology/state-laws-related-to-internet-privacy.aspx

[46] Karuna Joshi, Ontology for Data Privacy Policy, http://ebiquity.umbc.edu/resource/html/id/370/Ontology-for-Data-Privacy-Policy