A Semantic-based Approach to Reduce the Reading Time of Privacy Policies

by

Jasmin Kaur

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Computer Science

Guelph, Ontario, Canada

© Jasmin Kaur, January, 2020
ABSTRACT

A SEMANTIC BASED APPROACH TO REDUCE THE READING TIME OF PRIVACY POLICIES

Jasmin Kaur
University of Guelph, 2020

Advisor(s):
Dr. Rozita Dara
Dr. Charlie Obimbo

Privacy policy is a legal document in which the users are informed about the data practices used by the organizations. Past research indicates that the privacy policies are long and hard to understand. They are also known to have incomplete content. Users are not inclined to read the policy as they have to read long policies to find information about data practices of an organization. The solution that we are proposing in this research is to assist users with finding relevant content to their queries using semantic approach.

This thesis presents the development of domain ontology for privacy policies. Natural Language Processing was used to understand the content of the policies and capture vocabulary for the ontology. This vocabulary was further used to build the ontology so that the ontology highlights relevant sentences related to a privacy concern. We validated and evaluated the ontology using different methods: competency questions, data driven, metric based and user evaluation. Results from the evaluation of ontology show that the amount of text to read is significantly reduced as the users have to only read selected text that ranged from 1% to 30% of a privacy policy. The amount of text depended on the query and its associated keywords. This signifies that the time required to read a policy is significantly reduced as the ontology directs user to the right content for a query. This finding was also confirmed by the results of the user study session. The results from the user study session indicated that the users found ontology helpful in finding relevant selected sentences to read as compared to reading the entire policy.
ACKNOWLEDGEMENT

First, I would like to express my gratitude to my advisor Dr. Rozita Dara for her constant support and advice throughout my research. Her vast knowledge, guidance and patience has helped me get through my masters. Next, I would like to thank my co-advisor Dr. Charlie Obimbo for his inputs and continuous support during my research. His encouragement and generous advice has helped me throughout my masters.

Additionally, I would like to thank Dhiren Audich for providing the dataset of privacy policies and Curtis Rasmussen for providing the data set with the user concerns. I would also like to thank the participants who took their time to participate in the user study session.

Also, I am forever grateful to my parents and my brother for their love, support and encouragement. This would have not been possible without their help. Lastly, I would like to thank my friends and extended family who have helped me during my study in Canada.
# TABLE OF CONTENTS

Abstract ........................................................................................................................................... ii
Acknowledgement ....................................................................................................................... iii
Table of Contents ........................................................................................................................... iv
List of Tables .................................................................................................................................. viii
List of Figures ............................................................................................................................... ix
List of Abbreviations ..................................................................................................................... xi

Chapter 1  Introduction .................................................................................................................. 1
  1.1 Privacy Policies ...................................................................................................................... 2
  1.2 Ontology .................................................................................................................................. 3
    1.2.1 Components of Ontology ............................................................................................... 4
    1.2.2 Building Ontology ........................................................................................................ 4
    1.2.3 Ontology Evaluation ....................................................................................................... 5
  1.3 Evaluation of Ontology through User Studies ......................................................................... 6
    1.3.1 User Study Methods ...................................................................................................... 6
  1.4 Content Analysis of Privacy Polices ......................................................................................... 7
    1.4.1 Data Pre-processing ........................................................................................................ 7
    1.4.2 Topic modelling/Coverage ............................................................................................... 7
    1.4.3 Keyword Analysis .......................................................................................................... 8
    1.4.4 Similarity Measure ......................................................................................................... 8
  1.5 Methodology .......................................................................................................................... 8
    1.5.1 Problem Statement ......................................................................................................... 8
    1.5.2 Proposed Solution ......................................................................................................... 9
1.6 Contributions ........................................................................................................... 10
1.7 Thesis organization ................................................................................................. 10

Chapter 2 Content Analysis of Privacy Policies ............................................................... 12
  2.1 Experiments ........................................................................................................... 13
  2.2 Experimental Settings ........................................................................................... 14
    2.2.1 Privacy Policy Sections .................................................................................. 14
    2.2.2 Privacy Policy Corpus .................................................................................... 15
    2.2.3 Seed Keywords .............................................................................................. 16
    2.2.4 Ambiguous Keywords .................................................................................... 17
  2.3 Results ................................................................................................................... 18
    2.3.1 Topic Coverage analysis ............................................................................... 18
    2.3.2 Keyword analysis ........................................................................................... 21
    2.3.3 Ambiguous Keywords Analysis ..................................................................... 21
  2.4 Discussion .............................................................................................................. 22
    2.4.1 Data Protection Regulation Analysis ............................................................. 22
    2.4.2 Domain-specific Policy Analysis .................................................................. 24

Chapter 3 Ontology ........................................................................................................ 28
  3.1 Methodology .......................................................................................................... 28
    3.1.1 Purpose of Building Ontology ...................................................................... 28
    3.1.2 Proposed Approach ......................................................................................... 29
    3.1.3 Scope of Ontology .......................................................................................... 29
    3.1.4 Source of Ontology Development .................................................................. 30
    3.1.5 Building Ontology-Steps and Process .............................................................. 31
    3.1.6 Competency Questions .................................................................................... 31
3.2 Building an Ontology ........................................................................................................ 32

3.2.1 Structure of Ontology .................................................................................................. 32

3.2.2 Relations in Ontology .................................................................................................. 35

3.2.3 Competency Questions ............................................................................................... 36

3.3 Validation .......................................................................................................................... 38

3.4 Discussion ......................................................................................................................... 42

Chapter 4 Evaluation of Ontology using User Study ......................................................... 46

4.1 Task Based User Study ..................................................................................................... 46

4.2 Methodology .................................................................................................................... 47

4.2.1 Motivation and Objective .......................................................................................... 47

4.2.2 Proposed Approach .................................................................................................... 47

4.3 Experimental Setting ....................................................................................................... 48

4.3.1 Planning and Formulating Tasks .............................................................................. 48

4.3.2 Selecting Policies and Queries .................................................................................. 48

4.3.3 Questionnaire ............................................................................................................ 49

4.3.4 User Study Session Setup ........................................................................................ 50

4.3.5 Setting up User Study ............................................................................................... 50

4.4 User Study Metrics ......................................................................................................... 51

4.5 Results .............................................................................................................................. 52

4.6 Discussion ......................................................................................................................... 58

Chapter 5 Conclusion ............................................................................................................ 61

5.1 Limitations & Future Work ............................................................................................. 62

References .................................................................................................................................. 63

Appendices ................................................................................................................................ 69
LIST OF TABLES

Table 2.1 Seed Keywords and Data Privacy Sections .................................................. 17
Table 2.2 Ambiguous Words ....................................................................................... 18
Table 2.3 Coverage of Data Privacy Sections in Different Domain ......................... 20
Table 3.1 List of Competency Questions considered for building ontology ............ 32
Table 3.2 Example of a class in the ontology ................................................................ 33
Table 3.3 Object properties used in ontology along with the domain, range and characteristics ..................................................................................................................... 35
Table 3.4 Base metrics ................................................................................................. 39
Table 3.5 Schema metrics ........................................................................................... 40
Table 3.6 Knowledgebase metrics .............................................................................. 40
Table 3.7 Results for Experiment 1: Policy Coverage .............................................. 41
Table 3.8 Results for Experiment 2: Completeness .................................................... 42
Table 4.1 The selected privacy policy along with the query or privacy concern ....... 49
Table 4.2 Mean time per task ...................................................................................... 55
Table 4.3 Ease of use .................................................................................................. 56
LIST OF FIGURES

Figure 1.1 Overview of building and validating an ontology .................................................. 10
Figure 2.1 Overview of Content Analysis of Privacy Policies .................................................. 14
Figure 2.2 Coverage of Data Privacy Sections in 2000 Policies ............................................ 19
Figure 2.3 Coverage of Data Privacy Sections in USA, Canada and UK .................................. 20
Figure 3.1 Overview of how Ontology is built and validated ................................................... 30
Figure 3.2 Steps in building ontology ...................................................................................... 31
Figure 3.3 Classes in Ontology ............................................................................................... 34
Figure 3.4 Example of individuals for a class in ontology ....................................................... 34
Figure 3.5 Data properties and its data type used in the ontology .......................................... 36
Figure 3.6 Sentences selected for Q4: Does this policy have an opt in/opt out policy? ........... 44
Figure 3.7 Sentences selected for Q7: Does this policy discuss about data security practices? .................................................................................................................................................. 44
Figure 3.8 Sentences selected for Q11: Does this website share my personal information with third parties? .............................................................................................................................................. 45
Figure 4.1 Computer literacy and Internet literacy of the users .............................................. 52
Figure 4.2 Users concern for privacy protection and data practices ........................................ 53
Figure 4.3 Time taken by the users to complete task 2 and task 3 for Amazon policy .......... 54
Figure 4.4 Time taken by the users to complete task 2 and task 3 for Honda policy ............... 54
Figure 4.5 Time taken by the users to complete task 2 and task 3 for Google policy .......... 55
Figure 4.6 Efficiency of the tasks based on different policies ................................................ 56
Figure 4.7 Likert scale score for task 2 and task 3 based on expectation ................................. 57
Figure 4.8 Likert scale score for task 2 and task 3 based on effort to complete task .......... 58
Figure 4.9 Likert scale score for task 2 and task 3 based on time to find information .......... 58
Figure C.0.1 Sentences selected for Q1: Which PII are covered by this policy? ............ 76
Figure C.0.2 Sentences selected for Q2: Does this website track me? 76

Figure C.0.3 Sentences selected for Q3: Does this website use cookies? 77

Figure C.0.4 Sentences selected for Q5: Does this policy specify any data retention practices? 77

Figure C.0.5 Sentences selected for Q6: Will the user be notified of any changes to this policy? 77

Figure C.0.6 Sentences selected for Q8: What choices are available in the policy? 78

Figure C.0.7 Sentences selected for Q9: Can I have access to my data? 79

Figure C.0.8 Sentences selected for Q10: Which PII are being shared under this policy? 79

Figure C.0.9 Sentences selected for Q12: Does this policy discuss about tracking cookies? 80
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMU</td>
<td>Carnegie Mellon University</td>
</tr>
<tr>
<td>COPPA</td>
<td>Children’s Online Privacy Protection</td>
</tr>
<tr>
<td>CP</td>
<td>Concurrent Probing</td>
</tr>
<tr>
<td>CQ</td>
<td>Competency Question</td>
</tr>
<tr>
<td>CTA</td>
<td>Concurrent Think Aloud</td>
</tr>
<tr>
<td>DAML</td>
<td>DARPA Agent Markup Language</td>
</tr>
<tr>
<td>DL</td>
<td>Description Logic</td>
</tr>
<tr>
<td>EFF</td>
<td>Electronic Frontier Foundation</td>
</tr>
<tr>
<td>FIP</td>
<td>Fair Information Practices</td>
</tr>
<tr>
<td>FTC</td>
<td>Federal Trade Commission</td>
</tr>
<tr>
<td>GDPR</td>
<td>General Data Protection Regulation</td>
</tr>
<tr>
<td>HIPPA</td>
<td>Health Insurance Portability and Accountability Act</td>
</tr>
<tr>
<td>LDA</td>
<td>Latent Dirichlet Allocation</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>OECD</td>
<td>Organization for Economic Cooperation and Development</td>
</tr>
<tr>
<td>OIL</td>
<td>Ontology Inference Layer</td>
</tr>
<tr>
<td>OWL</td>
<td>Web Ontology Language</td>
</tr>
<tr>
<td>P3P</td>
<td>Platform for Privacy Preferences</td>
</tr>
<tr>
<td>PET</td>
<td>Privacy Enhancing Technologies</td>
</tr>
<tr>
<td>PII</td>
<td>Personally Identifiable Information</td>
</tr>
<tr>
<td>PIPEDA</td>
<td>Personal Information Protection and Electronic Documents Act</td>
</tr>
<tr>
<td>RDF</td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td>RP</td>
<td>Retrospective Probing</td>
</tr>
<tr>
<td>RTA</td>
<td>Retrospective Think Aloud</td>
</tr>
<tr>
<td>SEQ</td>
<td>Single Ease Question</td>
</tr>
<tr>
<td>SPARQL</td>
<td>SPARQL Protocol And RDF Query Language</td>
</tr>
<tr>
<td>SSL</td>
<td>Secure Socket Layer</td>
</tr>
<tr>
<td>TF</td>
<td>Term Frequency</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>USA</td>
<td>United States of America</td>
</tr>
<tr>
<td>W3C</td>
<td>World Wide Web Consortium</td>
</tr>
</tbody>
</table>
Chapter 1  Introduction

With the growing popularity of digital media, there is an increasing concern for privacy of the users as their personal information is collected and used. The collection of personal data is mostly intended to make users’ experience more enjoyable and productive. However, the excessive collection of personal data may lead to tracking of movements of users’ behavior, browsing history and buying patterns without their consent and awareness. Another concern is that users’ data can be shared and sold to third parties. Furthermore, user data can be exploited to steal or misuse person’s digital identity. Lastly, a major issue with most personal data collection practices is that users have limited knowledge and control over how their data is collected or even used [1]. This has resulted in an increased concern for privacy of the individuals.

Privacy policies are intended to help reduce some of these concerns [2] and make the users aware of the service providers’ data practices. Privacy Enhancing Technologies (PET) have been introduced to reduce the collection of personal data, use of opt out options, or anonymize personal data. However, lengthy privacy policies and use of elusive, vague terminology is still a problem. Due to these reasons, users do not read the entire policy and usually agree to it with limited awareness about the data practices. This problem can be solved if the users only have to read some sentences regarding a privacy concern instead of reviewing the entire policy.

Domain ontology can be helpful in structuring the concepts and describing relationships between concepts in the policies. With the help of domain ontology, we can capture the vocabulary and competency questions which can help the users to find the required information for a privacy concern in a lengthy policy. This also assists in better understanding and awareness of the data practices in a shorter period of time.

In this thesis, we present the development of domain ontology which captures the concept of privacy policies. To achieve this, we first have to analyze the content of privacy policies to determine the important vocabulary to be used in the ontology. Then, we will build the ontology and evaluate them using different approaches and metrics.

This chapter introduces the different topics related to the thesis which will help in understanding the objectives for our research. The rest of this chapter is organized as:

1. Section 1.1 provides an overview of privacy policies
2. Section 1.2 discusses the content analysis of privacy policies
3. Section 1.3 describes the introduction of ontology, building and evaluating ontology
4. Section 1.4 introduces to the user study and its methods
5. Section 1.5 covers the problem and discussed how it can be solved
1.1 Privacy Policies

The online service providers collect, store and share personal information of the users without the user’s explicit permission and control. This problem is solved by using privacy policies. Privacy policy is a legal document in which the users are informed about the data practices used by the companies. These documents are important in building trust by being transparent about why and how their personal information is collected and used [3] [4]. Privacy policies can also help govern and regulate the data related practices.

Privacy policies are written in natural language format. The privacy policies are often long and have inconsistent language. Furthermore, they have no standardized format for disclosing the information and the user has to read the entire policy to know about a certain privacy concern. The companies use ambiguous and legal terminology to describe the data practices in policies [5]. According to McDonald et al., the estimated time cost reading privacy policies is approximately 201 hours a year with the worth about $3,534 yearly per American internet user [6]. Privacy policies lack clarity and require reading skill considerably higher than an average literacy level [7] [8]. Milne et al. mentioned that one of the reasons users do not read policy is because the policies are long and are less comprehensible due to the legal terminology [13].

Some efforts have been made to help the users in ease of reading the policies. The Platform for Privacy Preferences (P3P) is a machine-readable format for online privacy policies. It is formally developed by the World Wide Web Consortium (W3C) and provides transparency in policies which can be used to improve privacy protection [9]. However, it lacks standardized and comprehensive semantics which makes it difficult in validating policies and creating accurate policies [10].

Guntamukkala et al. used machine learning algorithms to categorize different sections of privacy policies and to evaluate completeness of the content according to the fair practices. The authors used keywords extraction methods to annotate privacy policies [11]. The keywords were extracted from various acts and principles using Latent Dirichlet Allocation (LDA) and other keyword extraction techniques. In total, 1300 keywords were extracted and analyzed. Privacy practitioners participated in selecting the top 100 words. Then, these selected keywords were used to annotate privacy policies and train the machine learning algorithms.

Wilson et al. considered annotating the policies to analyze the structure and complexity of privacy policies. Approximately, 23k fine grained data practice annotations were extracted from 115 privacy policies. The authors categorized the data privacy sections into 12 categories and analyzed the distribution of data privacy sections per policy and found that the distribution of the data privacy section varies with policies [12]. Each policy took an average of 72 minutes to annotate which can be time consuming.

Crowdsourcing is an approach in which information is obtained in a task which is performed by online workers. Ramanath et al. used a combination of machine learning and crowdsourcing to semi-automate extraction of key aspects of privacy policies of
websites [14]. The main idea is to decompose crowdsourcing task of answering questions about privacy policies that does not require workers to read entire policies. The sections of the privacy policy which were relevant to a given question were only presented and took 45s per question.

Wilson et al. found that crowdsourcing privacy policy annotation has little importance as the crowd workers are not privacy experts and since the policies are vague, it makes the task more difficult. The authors found 80% crowd agreement threshold and highlighting sections increased efficiency as the task completion rate was decreased due to less reading [15].

Oltramari et al. used query based semantic analysis to understand privacy policies. The authors built PrivOnto, an ontology, using detailed annotations which were performed by domain experts and mapped 23,000 annotated data practices from 115 US policies [16]. In order to expand the ontology, a new policy has to be annotated which takes 72 minutes to annotate a single policy which is very time consuming. Also, PrivOnto is only limited to US privacy policies.

In order to reduce the cost and speed up the development of ontology, Audich et al [17] used Natural Language Processing (NLP) algorithms to build a domain ontology. The authors used supervised and unsupervised algorithms to gather vocabulary for the ontology. The vocabulary was further mapped to the privacy categories and queries (or the privacy concerns) which helped in reducing the textual reading for a privacy concern. However, the initial ontology has limited vocabulary. Therefore, extension of the ontology is required to expand the concepts and privacy concerns which will be covered in this thesis.

1.2 Ontology

An ontology is a formal explicit description of concepts, attributes and relationships of concepts in a domain. It is an important method in generating semantic information as the ontology provides domain knowledge in the form of interrelated concepts. Furthermore, ontology is used to define common representation of data from heterogenous database and link one piece of information to the other information. Traditionally, ontologies capture the domain specific taxonomy which can be useful in providing insights into how the keywords are related to each other[80]. An ontology along with the set of instances of classes represents a knowledgebase.

There are different views for defining an ontology. Gruber originally defined ontology as an “explicit specification of a conceptualization” [19]. Borst [20] defined an ontology as a “formal specification of a shared conceptualization”. Later on, Studer combined these definitions as “an ontology is a formal, explicit specification of shared conceptualization [21]”. Formal means that the ontology is machine readable. Explicit implies the concepts and constraints which are explicitly defined. Shared refers to the idea that the ontology captures shared knowledge to a group and not to certain private individuals.
Conceptualization means an abstract model of some phenomenon and its related relevant concept.

Sowa defined ontology as a study of the categories of things that belonged to a certain domain [22]. Parallelly, Guarino [23] referred an ontology as a set of logical axioms designed to account for the intended meaning of a vocabulary.

There can be many reasons for developing an ontology [24]. Firstly, it can be used to share the structure of information among people or software agents. Secondly, ontologies can help in reusing the domain knowledge. Several ontologies can be integrated or extended to represent a large domain knowledge. Thirdly, the domain assumptions are stated clearly and can help in making changes to the ontology. Lastly, ontology is helpful in analyzing domain knowledge.

### 1.2.1 Components of Ontology

Some of the common components of ontology are listed below [25]:

1. **Classes**: It is the collection of objects or individuals and is also called concepts.
2. **Individuals**: Individuals are instances of concepts or objects. It represents the atomic level of ontology
3. **Attributes**: the concepts or classes have some properties or features which are described by a set of attributes
4. **Relations**: It describes how the concepts are related to each other. Relations can have specific properties such as symmetry, transitivity etc.
5. **Axioms**: Assertions in a logical form that comprise the overall theory that the ontology describes in its domain of application.

### 1.2.2 Building Ontology

There are many ways to model a domain which is heavily dependent on the application and use. Developing an ontology is an iterative process. To build an ontology, firstly, the domain and scope of the ontology is determined [24]. It includes determining the domain of the ontology, its use and the questions which the ontology would answer with the domain information. There are some general steps that are considered while building an ontology which are listed below [24][25]:

1. List important terms in the ontology
2. Define the classes and class hierarchy
3. Define relations
4. Define attributes
5. Define instances
6. Define axioms, rules, functions

A comprehensive list of terms is useful in representing concepts, relationships among terms and properties of concepts. The next step, defining the classes and class hierarchy, can be done using different approaches: top-down, bottom-up and the combination of both approaches. The use of approach can depend on the developer and the domain. Defining the other components of ontology such as relations, axioms, attributes and instances help in enriching the knowledge representation in the ontology.

1.2.3 Ontology Evaluation

Brank et al. discussed various levels of ontology: lexical/vocabulary/concept/data, hierarchy/taxonomy, semantic relations, context/application, syntactic, structure/architecture/design [26]. Based on these levels of ontology, Brank et al. discussed about the approaches of evaluation. The common approaches to evaluate ontology are based on application, data driven, golden standard, assessment by humans. The emphasis is on the fact that there is not any preferred approach to evaluate ontologies as it depends on the purpose of evaluation and the application.

- **Application based evaluation:** In this approach, the ontology will be used in some application. It is a simple way to evaluate ontology by plugging the ontology in the application and evaluating the results.
- **Golden standard:** The evaluation is performed by comparing an ontology to a “golden standard” which may also be an ontology. Prepping for the golden standard can be a lot of work, however, once the golden standard is outlined, it can automatically perform the evaluation.
- **Data driven evaluation:** The ontology is evaluated by comparing it with the data i.e. the concepts or specific terms from the domain that the ontology refers.
- **Assessment by human experts:** Evaluation can be performed manually by human experts based on decision criteria. They manually access how well the ontology meets the necessary requirements or standards.
- **Competency questions**: The competency questions are the requirements that are in the form of questions which the ontology is intended to answer. The questions are evaluated based on the completeness theorem [76].

- **Metric-based evaluation**: The quality of an ontology can be evaluated empirically using metrics which can evaluate the structure, schema and the knowledge-base [77].

### 1.3 Evaluation of Ontology through User Studies

The ontology is further evaluated by the users by conducting a user study session. User study focuses on observing the users’ behavior, attitude, and need by using observational techniques, task analysis and feedback methods. Generally, user study is used to identify any problems or gaps in a product or system and identify whether the system is usable or not and the degree of user friendliness.

According to Christopher, Robert Kosara, et al., conducting user study can help in identifying which method can be appropriate for a given situation. Secondly, it can be useful in getting to know why a particular method is effective [27].

#### 1.3.1 User Study Methods

1. **Task analysis**: It is a process in which users are given a task to perform and are observed how they perform their task [28].

2. **Contextual interview**: In contextual interview, the researcher listens or watch the user working in his natural environment. They are usually more natural and realistic and doesn’t usually have any tasks or scenarios [29].

3. **Focus groups**: It is a moderated discussion with 5 to 10 participants to know about the users’ beliefs, attitude and ideas [30].

4. **Individual interviews**: In individual interviews, the users are interviewed typically for 30 minutes to an hour. It helps to know about the users’ attitude, experiences and beliefs [31].

5. **Online surveys**: It is a structured questionnaire in which the user generally completes the form via internet. The online survey can vary in length and format. It helps to collect information with a very little cost [32].
1.4 Content Analysis of Privacy Policies

In order to better understand privacy policies, we are interested to perform content analysis of privacy policies. With this, we aim to provide a more in-depth insight into the content covered by privacy policies and, also, the most frequently used vocabularies and terminologies. We also intend to compare the content of privacy policies among different application domains and data protection regimes. Since privacy policies are written in natural language, we used Natural language processing (NLP) methods to pre-process and clean the policy documents for further analysis which is discussed in the following sub-section.

1.4.1 Data Pre-processing

NLP methods are often used to analyze text data. In NLP, data pre-processing is an important step as the methods of gathering data are loosely controlled. NLP helps in analyzing massive amount of text and helps find relevant information [18]. NLP consists of various steps and methodologies. We used the following in the experiments:

- **Tokenization**: Tokenization is the process of segmenting text into words or phrases which are called tokens. These words or tokens are further used to analyze text. In the process of tokenization, basic cleaning of data such as changing to lowercase and removing punctuation marks is performed [18].

- **Stop word removal**: The words which are not useful for analysis are called stop words. Some of these words can be verbs, conjunctions, and organization names. When these words do not carry important information for the analysis task in hand, they are usually removed. To remove them, a list containing stop words is formed so that they can be found and removed from the corpus as the text is being processed [18].

- **Stemming**: Stemming is the process of condensing a word to its root or word stem. For example, ‘collection’, ‘collects’, ‘collecting’ are the common representation of ‘collect’. With stemming, we are left with just the base words and redundancy is removed to improve results [18].

1.4.2 Topic modelling/Coverage

Topic modeling helps organize and extract topics from large textual data. It helps identify hidden patterns of words and categorize the documents based on topics or sections. We have considered using topic modeling technique to analyze the distribution of topics in the privacy policies. We used LDA for the topic modeling task.

**Latent Dirichlet Allocation**: LDA is a generative probabilistic method which is commonly used for topic modeling. Each document is considered to have a mixture of topics and each topic have a mixture of words. In LDA, the topic distribution is assumed to have a sparse Dirichlet prior that covers only a small set of topics and the topic is characterized by a set of words. This helps in better disambiguation of words and better assignment of
document to a topic. The topic is identified by automatically identifying the likelihood of term co-occurrence. To handle large amount of text, LDA treats documents as probability distribution of topics. LDA model is a three-level hierarchical Bayesian model and the documents can be associated to multiple topics. The LDA model is modular and can be extended to model relations between topics (Blei, Ng, & Jordan, 2003).

1.4.3 Keyword Analysis

Keywords analysis is useful as the keywords concisely represents the content of the document. As discussed in Section 1.2.2, the first step of building ontology is identifying the terms. We performed keyword analysis for the privacy policies to use the frequent keywords in building the ontology. We used Term Frequency (TF) for analyzing keywords:

**Term Frequency (TF):** Term Frequency is the raw count of the words in the document. It represents the counts of the number of times a word $t$ appears in a document $d$. If the raw count is denoted with $f_{t,d}$, then $tf$ can be represented as $tf(t,d) = f_{t,d}$. Other variations to $tf$ is Boolean frequencies, and logarithmically scaled frequency. The weight of the word in a document is proportional to the term frequency.

1.4.4 Similarity Measure

We were interested to know how similar the privacy policies are in terms of the content. For our experiments, we used Jaccard similarity to measure content overlap of the privacy policies. Jaccard similarity is used to measure the similarity and diversity of sample sets. To this end, we once used all keywords to measure similarity and then compared the content with only using seed keywords.

**Jaccard similarity** or intersection over union is a statistical method used to measure the similarity of sample sets such as textual data. It measures the diversity of finite sample sets, keywords in this study, and is defined as the size of the intersection of sets divided by the size of the union of sets. Mathematically, it is computed by:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

where $0 \leq J(A,B) \leq 1$.

1.5 Methodology

1.5.1 Problem Statement

As discussed before, privacy policies are long and have legal terminology which makes it difficult for an average user to read the whole privacy policy. Furthermore, privacy policies require reading skill higher than the average literacy level because of the legal
terminology and verbosity. Due to these reasons the users do not read the entire policy, hence, are not aware of data practices in the policy.

There have been some previous attempts in making privacy policies more usable and understandable. Machine learning methods and crowdsourcing have been used to better understand the policies and to evaluate the transparency of privacy policies. However, a gap still exists which is related to the amount of text the users have to read. An approach is required where the policy is highlighted based on the privacy concerns or data practices so that the readers can read specific text based on their concern.

1.5.2 Proposed Solution

Our proposed approach is to semantically analyze the policies to build an ontology which captures the domain knowledge which can help users by directing relevant content related to their privacy concern. We build an ontology according to the data practices and privacy policy domain knowledge. To achieve this, we first need to understand the content of privacy policies which can be performed by Natural Language Processing techniques. The content of policies is analyzed by performing keyword analysis and topic coverage. By analyzing the content of privacy policies, we build an ontology which captures the vocabulary, privacy concepts along with the competency questions. This helps in directing the users to the relevant text which caters to the privacy concern. We evaluate the ontology using different approaches. User study sessions are conducted to collect the users' views, and experience.
1.6 Contributions

The primary contributions of this research are as follows:

1. In order to build a domain ontology, understanding the content of privacy policies is important. This is achieved by performing keyword analysis and topic coverage on a varied dataset of privacy policies including different countries and application domain.

2. Building an ontology which captures the data practices and privacy concepts that would return keywords related to privacy concerns which would help in directing users to relevant sections related to the concern.

3. Conducting user study session to evaluate the ontology output and the user experience.

1.7 Thesis organization

Figure 1.1 shows the general steps to build and validate the ontology for privacy policies. The rest of the thesis is organized as follows:
• Chapter 2 presents the content analysis of privacy policies. Results from term frequency and similarity measure are discussed and was used to build the taxonomy for the ontology.

• Chapter 3 presents the domain ontology for privacy policies. The ontology is evaluated and validated using different approaches: using competency questions, CMU’s OPP115 and metric based evaluation.

• In Chapter 4, we evaluated the ontology by users. User study was conducted in which the users performed some tasks to evaluate the ontology.

• Chapter 5 discusses the results along with some limitations which can be done in the future.
Chapter 2  Content Analysis of Privacy Policies

As discussed in Section 1.1, privacy policies have inconsistent and incomplete content. One of the reasons for inconsistent privacy policies is the lack of standardization in terms of the content, structure and organization across various organizations and applications [33]. Due to this asymmetric information, the users are not bothered to read the entire privacy policy [34]. Furthermore, some studies indicate that to understand the policies, college reading level is required [8] [35]. To better understand the privacy policies and its sections, we performed a comprehensive content analysis to make useful inference from policies [36].

Content analysis is a grounded theory which can help in making inferences from the text data [37]. With content analysis, the keywords of the privacy policies can be analyzed to understand the content of privacy policies. To the best of our knowledge, no quantitative analysis of privacy policy keywords has been performed. Additionally, an analysis of the content of privacy policies for different application domains (e.g. health, e-commerce, children, and others) and Fair Information Practices (FIPs) (e.g. countries) does not exist.

However, some work has been performed to examine whether the policies adhere to the laws and regulations. Ryker et al. [38] found that only 5.7% policies comply with all the recommended fair information practices, 62.8% comply partially and 31.4% business related privacy policies failed to comply with one or more fair information practices. In a more recent study by Salva et al., the authors analyzed the content of health-related policies to evaluate readability and to check if the policies are in compliance with the privacy principles. The findings indicated that from 35 policies, only 26% of them were fully in compliance with FIP [39].

Stamey et al used latent semantic analysis to analyze e-commerce privacy policies to identify the important privacy sections and ranked the significant words for the identified sections [40]. The authors analyzed the semantic relations between words in the policies and identified ambiguity in text.

Machine learning algorithms have also been used to analyze the privacy related text documents. Liu et al. analyzed the privacy policies by using machine learning algorithms to segment and classify the policies into 12 data privacy sections. The authors used unsupervised labelling methods to analyze and label privacy policies automatically [41]. Similarly, Costante et al. proposed a supervised learning method to analyze data privacy sections covered in the privacy policies. Their results proved that using an automatic classifier can help effectively associate the appropriate category to paragraph of the policy document [42].

Guntamukkala et al. [11] used machine learning algorithms to categorize different sections of privacy policies and to evaluate completeness of the content according to the fair practices. The authors used keywords extraction methods to annotate privacy policies. The keywords were extracted from various acts and principles using LDA and other keyword extraction techniques. In total, 1300 keywords were extracted and analyzed.
Privacy practitioners participated in selecting the top 100 words. Then these selected keywords were used to annotate privacy policies and to train the machine learning algorithms.

In this chapter, we aim to provide a more in-depth insight into the content covered by privacy policies and, also, the most frequently used vocabularies and terminologies. We also intend to compare the content of privacy policies among different application domains and data protection regimes (FIPs).

Analysis of online privacy policies was performed using NLP methods and was focused on the following objectives:

- A comprehensive keywords and topic analysis
- Comparison of (a) commonly used keywords and ambiguous words and (b) the coverage of FIPs recommended sections in several application domains and data protection regulations.
- Quantitative analysis of keyword similarity among different application domains and data protection regimes.

We believe a comprehensive analysis of keywords, sections' coverage, diversity of keywords, and practices in using ambiguous words has several advantages:

- It provides a more in-depth insight into the coverage of different sections in policies. It also advances our understanding of how policies are written in accordance to the best practices and regulatory requirements.
- Analysis of policies related to different application domains can help us understand whether laws and acts impact the language and word choices in policies.
- Analysis of policies in the context of different data protection legislations can be insightful to understand how regulations impact the content of policies.

2.1 Experiments

We used NLP methods to pre-process and clean the policy documents for further analysis. An overview of the experiments is given in Figure 2.1. We performed the following three experiments:

- **Experiment 1**: We used a set of seed keywords, keywords that were extracted by Guntamukkala et al. [11] and Wilson et al., 2016 [12] for each section of privacy, to perform topic modeling and compare the content of policies. Topic modeling helps organize and understand the content of privacy policies and also identify patterns of keywords/terminology usage in the documents.
- **Experiment 2**: Use of common keywords, seed keywords, and ambiguous words in privacy policies were evaluated. For the first two, we calculated the frequency of words, either the seed keywords or the common keywords, and compared them among different privacy corpuses and regulations. For the ambiguous words, we first collected the most common words from the existing literature. We
then examined the trend of using those words in different regulatory regimes, applications and overall for all the privacy policies that we had collected.

![Figure 2.1 Overview of Content Analysis of Privacy Policies](image)

- **Experiment 3**: Overlap of the content of privacy policies was also investigated. After initial cleaning, we used a similarity measure to check the union of keywords between different privacy policy corpuses. We also compared the practice of using seed keywords among different policy sets.

### 2.2 Experimental Settings

#### 2.2.1 Privacy Policy Sections

Privacy policy sections that we considered in this study are originated from Federal Trade Commission (FTC), Organization for Economic Cooperation and Development (OECD) and FIP practices as well as recent work of Guntamukkala et al. [11] and Wilson et al. [12]. Since these FIP practices were influential in the creation and evolution of the content of privacy policies, we decided to use them as a framework to conduct our investigation. Furthermore, we combined the keywords that Guntamukkala et al. [11] extracted for each section of the policy with Wilson et al. [12] keywords. To do this, we had to aggregate privacy policy sections and their definitions in [11] and [12]. Privacy policy sections that we examined for our analysis are:

- **Collection**: this section explains what information is collected from the users and how is the information gathered.
- **Sharing**: this section states with whom (third party) and under what
circumstances the users’ information is shared.

- Choice: privacy related choices such as option, opt in, opt out comes under this section.
- Access: this section focuses on how a user can access and check accuracy of their personal information.
- Data retention: the process of retaining personal information of the users is stated in this section. It can basically cover the purpose, reason and duration of retention of users’ personal information.
- Data security: this section discusses standard practices of protecting users' information by the service provider.
- Policy change: it explains how the users will be informed about the change in privacy practices.
- Do not track: this section focuses on online tracking such as use of cookie, and other web technology which keeps track of the users’ data.
- Purpose: this section explains the purpose of collecting and using information by the service provider. The company should confirm that the data will not be used for any other purpose stated.

### 2.2.2 Privacy Policy Corpus

To perform the experiments, we first collected large sets of privacy policies to perform keyword analysis, and similarity measure and topic coverage in policies. It is important to note that we limited our study to privacy policies in English. We created three corpora of policies from various sources:

- **2000 policies**: We used 2000 policies that were collected by Massey et al. [43]. The authors collected these policies from different sources such as Google top 1000, Fortune 500, Health Insurance Portability and Accountability Act (HIPPA) health policies, and financial policies. Google top 1000 policies were collected based on the Google’s estimate of Internet traffic. Fortune 500 comprised of policies of the most influential companies. Our motive to collect 2000 policies is to analyze large number of policies irrespective of any domain, and geographic area. There might be chances
of policy overlap with the other corpora of policies.

- **600 policies**: We collected 600 policies covering 11 different application domains from Alexa.com [44]. The motivation is to analyze application domain specific policies only. However, there might be policy overlap with other dataset policies. The application domains are listed below:
  
  o Adult: policies related to adult finder
  o Business: policies related to financial banks, e-commerce, and business companies
  o Computers: policies related to technology, tech applications/software and tech companies
  o Games: policies related to various gaming sites and companies
  o Health: policies related to health, fitness and wellness
  o Kids: policies related to kids such as learning websites and gaming websites
  o Recreation: policies related to restaurants, travelling, airlines, and other recreational activities
  o Reference: policies related to learning, online courses, and other educational sites
  o Science: policies related to scientific work and educational/informative sites
  o Shopping: policies related to e-commerce and shopping online
  o Sports: policies related to sport news, and other sports related sites

- **50 policies**: We collected financial and e-commerce related privacy policies from United States of America (15 policies), Canada (15 policies) and Europe (5 German and 15 United Kingdom policies). The reason for considering only financial and e-commerce policies is because we are only interested in analyzing authentic regional policies.

### 2.2.3 Seed Keywords

We considered a set of keywords, called seed keywords, that were categorized based on the privacy policy sections [78]. Table 2.1 includes the set of seed keywords that were used in our experiments.
Table 2.1 Seed Keywords and Data Privacy Sections

<table>
<thead>
<tr>
<th>Sections</th>
<th>Seed Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collect</td>
<td>collect, personal, identifiable, telephone, number, phone, telephone number, IP address, phone number, IP, mobile, email, address, name, date of birth, birthday, age, account, credit card, location, username, password, contact, zip code, postal code, mailing address, phone</td>
</tr>
<tr>
<td>Share</td>
<td>party, share, sell, disclose, company, advertiser, provider, partner, public, analytics, companies, organizations, businesses, contractors, divulge, sell, law, legal, regulation, third party, transfer, service providers, marketing partners, subsidiaries, disclosure, safe harbor</td>
</tr>
<tr>
<td>Purpose</td>
<td>ads, use, services, verifying, purpose, fraud, prevention, improve products, identification, promotions, personalize, advertising, analytics</td>
</tr>
<tr>
<td>Choice</td>
<td>opt, unsubscribe, disable, choose, choice, consent, setting, option, wish, agree, opt-in, opt-out, subscribe, do not track</td>
</tr>
<tr>
<td>Access</td>
<td>delete, profile, correct, account, change, update, section, access, removal, request, modify, edit, settings, preferences, accurate</td>
</tr>
<tr>
<td>Retention</td>
<td>retain, store, delete, deletion, database, participate, promotion, send friend, record, remove, retention, keep, data, backup, discard</td>
</tr>
<tr>
<td>Security</td>
<td>secure, security, safeguard, protect, compromise, encrypt, advertiser set, unauthorized, access, SSL, socket, socket layer, encryption, restrict, fraud</td>
</tr>
<tr>
<td>Policy Change</td>
<td>change, change privacy, policy time, current, policy agreement, update privacy, update, notice</td>
</tr>
</tbody>
</table>

2.2.4 Ambiguous Keywords

We also examined the use of ambiguous words in privacy policies. The list of words is shown in Table 2.2. This list is taken from Guntamukkala et al. [11] and Reidenberg et al. [45].
2.3 Results

2.3.1 Topic Coverage analysis

We analyzed the coverage of each of the recommended sections in privacy policies using LDA. Figure 2.2 depicts the coverage of nine sections in 2000 policies. The results suggest that the coverage of collection section is the highest with 23.45%. This was expected as collection has the largest number of seed keywords and covers different data practices. Despite our expectation, sharing section did not have a high coverage. The coverage of choice section is the lowest (5.49%), following data retention (5.65%), and do not track (5.52%), which suggests that the privacy policies give little attention to informing users about their choices to set their privacy preferences and, also, about the service providers’ tracking or retention practices.
We also measured the coverage of privacy sections among different regulations. Figure 2.3 represents the coverage for USA, Canada, and Europe. Similar to the 2000 privacy policy corpus, collection has the highest coverage. Other sections follow very similar patterns for coverage as they did in 2000 privacy policy corpus. The two notable differences are in the do not track and retention sections in European policies and access in USA policies. Purpose has the highest coverage in USA policies.

Table 2.3 shows the coverage of privacy sections in policies of different application domains. One interesting finding is that, although healthcare is considered to be one of the most sensitive application domains, access and retention sections are not extensively covered. However, security is given more attention than other domains. On the other hand, gaming domain pays more attention to retention and do not track topics. In addition, policies related to kids have a higher coverage for purpose, sharing, and do not track. A consistent observation, regardless of the size of privacy corpus, domain, or regulatory framework is that policy change is not covered extensively in policies. This is despite the fact that notice and user consent are highly regarded and are recommended privacy practices.
### Table 2.3 Coverage of Data Privacy Sections in Different Domain

<table>
<thead>
<tr>
<th>Sections</th>
<th>Adult</th>
<th>Business</th>
<th>Computers Games</th>
<th>Health</th>
<th>Kids</th>
<th>Recreation</th>
<th>Reference</th>
<th>Science</th>
<th>Shopping</th>
<th>Sports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection</td>
<td>27.9%</td>
<td>25.7%</td>
<td>27.6%</td>
<td>26.2%</td>
<td>24.4%</td>
<td>02.5%</td>
<td>25.9%</td>
<td>22.9%</td>
<td>22.9%</td>
<td>24.9%</td>
</tr>
<tr>
<td>Sharing</td>
<td>16.3%</td>
<td>16.6%</td>
<td>17.9%</td>
<td>14.8%</td>
<td>14.3%</td>
<td>16.7%</td>
<td>14.3%</td>
<td>13.3%</td>
<td>15.3%</td>
<td>14.5%</td>
</tr>
<tr>
<td>Purpose</td>
<td>14.5%</td>
<td>19.17%</td>
<td>15.4%</td>
<td>15.8%</td>
<td>19.7%</td>
<td>21.5%</td>
<td>15.9%</td>
<td>17.3%</td>
<td>18.6%</td>
<td>17.3%</td>
</tr>
<tr>
<td>Choice</td>
<td>7.5%</td>
<td>6.82%</td>
<td>6.5%</td>
<td>4.4%</td>
<td>05.5%</td>
<td>04.4%</td>
<td>05.8%</td>
<td>08.4%</td>
<td>05.0%</td>
<td>07.7%</td>
</tr>
<tr>
<td>Access</td>
<td>5.8%</td>
<td>5.61%</td>
<td>6.6%</td>
<td>6.9%</td>
<td>07.9%</td>
<td>07.0%</td>
<td>10.6%</td>
<td>07.4%</td>
<td>07.1%</td>
<td>12.3%</td>
</tr>
<tr>
<td>Retention</td>
<td>8.3%</td>
<td>7.2%</td>
<td>6.6%</td>
<td>8.8%</td>
<td>07.2%</td>
<td>05.5%</td>
<td>08.6%</td>
<td>07.8%</td>
<td>07.2%</td>
<td>07.2%</td>
</tr>
<tr>
<td>Security</td>
<td>8.2%</td>
<td>6.9%</td>
<td>7.7%</td>
<td>5.8%</td>
<td>08.0%</td>
<td>05.8%</td>
<td>05.1%</td>
<td>05.6%</td>
<td>06.1%</td>
<td>06.9%</td>
</tr>
<tr>
<td>Policy Change</td>
<td>4.3%</td>
<td>3.6%</td>
<td>3.2%</td>
<td>2.9%</td>
<td>03.5%</td>
<td>04.5%</td>
<td>03.5%</td>
<td>05.6%</td>
<td>03.3%</td>
<td>03.7%</td>
</tr>
<tr>
<td>Do Not Track</td>
<td>6.8%</td>
<td>8.2%</td>
<td>7.9%</td>
<td><strong>12.7%</strong></td>
<td>07.4%</td>
<td>09.8%</td>
<td>10.3%</td>
<td>09.8%</td>
<td>12.6%</td>
<td>08.8%</td>
</tr>
</tbody>
</table>

Figure 2.3 Coverage of Data Privacy Sections in USA, Canada and UK.
2.3.2 Keyword analysis

We also analyzed the frequency of words after pre-processing of 2000 privacy policies using stop-word removal and stemming. Most frequent terms were:

- personal, information, may, terms, third party, content, access, email, account, rights, agree, shall, contact, address, collect, notice, business, company, security, certain, share, purpose, name, number, customer, identifiable, advertising, trademarks, cookies, must, consent, browser, change, credit, protect

Among these terms, we can recognize several seed keywords which are highlighted in bold and a few ambiguous words which are underlined. We can observe that words such as third party, access, email, cookies, address, and a few other Personally Identifiable Information (PII) were frequently used. We also compared frequent words in European policies with the 2000 policy corpus. Cookies was on the top in European policies, followed by purpose, collect, and IP. Secure and consent were two of the top keywords in Canadian policies. Frequency of words in US policies was similar to the 2000 policy corpus. An interesting observation was that share and disclose are not used in German policies in comparison to other regions policies. This could be due to the fact that German laws have enforced limitations on personal data sharing and disclosure and such data protection practices are embedded in their regulations. Lastly, an analysis of seed keywords in different application domains suggested that shopping policies mention PII such as name, email, address, phone, mobile, contact, mail, location more frequently in comparison to other domains. The results of the similarity measure are given in Appendix A.

2.3.3 Ambiguous Keywords Analysis

Ambiguity is subjective and cannot be accurately measured with just examining frequency of words in a text. One needs to investigate the context in which such words are used. Also, a sentence can be ambiguous without using words that are highlighted in Table 2. However, frequency of ambiguous words in a text can be an indication of ambiguous content. The more ambiguous words are used, the more a text may lack transparency and clarity. Therefore, we believe that users can be warned about the ambiguous content by examining the ambiguous words used in the policies.

Examining frequency of ambiguous words in the 2000 privacy policy corpus depicted that the most frequent words were may, certain, for example, except, and like. Furthermore, it is observed that the top three ambiguous words in USA, Canada, UK and German policies are may, will, and can. Additionally, Canadian policies made use of appropriate frequently, while German policies used generally and like frequently.

We also calculated the percentage of lines that included an ambiguous word in each policy corpus. For the 2000 privacy policy corpus and almost all application domains,
The top 5 ambiguous words in 11 different domains are nonetheless the same. These are: *may, share, certain, like, and for example*. Computers domain used the ambiguous words most frequently followed by shopping domain. Moreover, the largest number of ambiguous lines was reported in the recreation domain and the smallest number of ambiguous lines was in the kids’ domain. Overall, in all the domains, *may* was the most frequently used ambiguous words.

### 2.4 Discussion

We further investigated the impact of privacy regulations on privacy policy language choices. To achieve this, we selected several data practices in each region and examined the topic coverage and pattern of seed keyword usage related to those data practices. Our findings are discussed in the following sections.

#### 2.4.1 Data Protection Regulation Analysis

**United Kingdom**

According to UK’s Data Protection Act 1998, collection, storage and use of personal data is permitted. However, it should be within the constraints of the law. For example, requiring *email* and *name* to access a website is not a recommended practice [46]. We noticed that PII such as *name, email, and address* are used often in the UK policies. This signifies that the companies may in fact be collecting these PIIs. We further looked into some UK policies to check how the collection practices are described. For example, the following lines are taken from Made policy [47]:

*We describe the categories of personal information we collect in more detail below:*

(a) **Contact details:** such as your name, your email address, your telephone number and addresses associated with your account.

(b) **Profile information:** such as your contact details (as above) and other any information you share when creating a profile on Unboxed.

These statements as well as additional data collection statements in this policy suggest that Made does not comply with UK data collection limitation regulation in particular in relation to PII data. We also observed that *access* is used only in 5.9% of UK policies that we have collected. This score represents the percentage of sentences that used related keywords in the access category (Table 2.3). We noted that under the right of *subject access* in the Data Protection Act, the individuals have the right to access or ask for their personal information held by the companies [48]. The following sentences were taken
from Barclays privacy policy as an example to depict how some service providers comply with this right [49]:

_The Data Protection Act 1998 gives you the right to access information held about you. Your right of access can be exercised in accordance with the Act. Any access request may be subject to a fee of £10 to meet our costs in providing you with details of the information we hold about you._

According to this policy, individuals have to pay to access to their personal information.

Similar to other countries, UK law also regulates the principle of data retention [50]. These regulations require that policies to be transparent about what type of data is retained, cost of retention and for how long the data is retained. Despite this requirement, we observed that seed keywords such as _retention, retain and record_ were not used as often as compared to policies of other countries. Some policies even did not have a section dedicated to retention.

**Canada:**

According to 2016 Survey of Canadians on Privacy conducted by OPC, majority of people (86%) feel that websites should get consent before using their information to create targeted online advertisements [51]. This is reflected in the Canadian policies as well. According to PIPEDA, companies are required to get meaningful consent for collection, disclosure and use of personal information [52]. Interestingly, we noticed that the keyword _consent_ is used in 95% of the Canadian policies. We further examined how consent is used in the policies and in what context. The following lines are extracted from the Indigo privacy policy [53]:

_We will collect, use or disclose your personal information only with your knowledge and consent, except where required or permitted by law._

* … will not make your consent a requirement to the supply of a product or a service other than required to be able to supply the product or service.*

* You may withdraw your consent at any time, on reasonable notice, subject to legal or contractual restrictions. …*

As it can be observed in the sample policies above, the word _consent_ is not necessary used in the context of the right of users. It is used to explain the process in which users consent to. Another interesting observation was that _access_ is used frequently in Canadian policies in comparison to other countries. According to the Canadian law, access to data is the basic right of an individual which should be facilitated. Other most frequent seed keywords are _update_ and _delete_ in the Canadian policies. This suggests that users can have access to their data and are given a choice to change their data.

_Cookie_ is used less in the Canadian privacy policies in comparison to other regions. This is despite the fact that a great deal of emphasis is made on the importance of informed
consent when it comes to collection of cookies in Canadian regulation. On the other hand, the use of security seed keywords is most common in the Canadian privacy policies.

**USA:**

USA laws emphasize on transparency in retention of data and the impact of this data practice is obvious in the policies. Seed keywords such as *retention* and *retain* are highly frequent in their policies. In fact, these words appear in all USA policies that we have collected. We also reviewed several policies and noticed that the words *retain*, and *retention* are scattered throughout the policy. The other most frequent words in USA policies are *notice, password,* and *SSL.*

In general, USA laws do not provide the right to access specific personal information, however, there are exceptions in some laws including HIPPA [55]. From our analysis, it is observed that *access* is used fairly frequently in health policies. Similarly, updating and deleting data is not common under the USA laws but this is an exception for health-related policies in which users have the right to update and delete data [55]. We have noticed that *delete* and *update* are not frequently used in USA policies. This could potentially mean that the policies are not transparent about access data practices.

USA laws do not require consent to use cookies and other web technology for tracking [54]. However, keywords such as *consent* and *cookies* are used frequently to explain the organization data practices for cookies. In addition, seed keywords related to tracking such as *third party, track, beacon* are used the more frequently in USA policies. The use of PII keywords such as *email, location,* and *phone* are also notable as compared to policies of other countries.

**Germany:**

It is interesting to note that seed keywords such as *phone, share, retain, disclose* and *password* are not used in the policies at all. Several other seed keywords also occurred less frequently such as *delete, access, email, update, retention,* and *notice.* Additionally, German laws have enforced limitations on their data collection practices and such practices are embedded deeply in their regulations. This fact has clearly had an impact on keyword choices that are used in privacy policies.

**2.4.2 Domain-specific Policy Analysis**

**Healthcare:**

Medical records can include personal information about a patient such as details of patient’s physical and mental health, personal relationships and financial information. This is reflected in the use of keywords in health-related policies since words such as *name, age,* and *address* are frequently used. The following lines are taken from the privacy policy which describe what personal information are asked from the users [56]:

---

24
“You will be asked to provide MyFitnessPal certain information related to You ("Personal Information"). Personal Information includes, without limitation: (1) “Contact Data” (personally identifiable information about You, such as Your name and email address, as well as Your friends and contacts, if you enable access to Your contacts and address book information); and (2) “Demographic Data” (personal information about You (but doesn’t specifically identify You), such as Your gender, birthday, zip code, country, height, weight, lifestyle and exercise frequency); and (3) “Fitness Data” (information about Your use of the Services (but doesn’t specifically identify You), such as Your caloric intake, nutritional statistics, fitness activity, and weight loss/gain).”

Since most of the health care companies gather highly sensitive personal information such health plan number and social security number, it was expected that a significant emphasis is given to “security” of personal health information in their policies. However, seed keywords related to security such as safeguard, security, and SSL are not used frequently. Furthermore, the other important aspect is access of medical information. Surprisingly, the coverage of access is only 7.9%. These privacy policies largely covered the collection and purpose. However, policy change, another important section in health-related policies, has the lowest coverage of 3.5%. The opt out option is not discussed frequently either.

Shopping:

The most common data privacy section in the shopping domain is collection and behavioural tracking especially use of cookies [54]. In our analysis of coverage of data privacy sections, it is noticed that coverage of collection section is relatively larger than the other sections. Similarly, sharing section also has a high coverage (15.3%). The following lines are taken from the Nike privacy policy which describe the sharing practice [57]:

“When you join or use certain services, you agree to publicly share a basic amount of information, which may include your username, city location, and profile picture.”

Additionally, this signifies that the most common and obvious data privacy section in the shopping domain is collection of personal information such as payment information, site analytics and behavioural tracking. This information is claimed to be collected to make the shopping experience more user friendly. From our experiments, we noticed that the seed keywords such as name, email, address, and contact information such as phone number, mobile, location were frequently used in e-commerce policies. Furthermore, username and password are also the most frequent words used in the e-commerce privacy policies.

For shopping, choice of opt-in or opt-out is critically important. However, choice is covered only 5% in this group of policies. Another important aspect is securing personal information which covers 6.9% of the policy. By further examining the e-commerce
policies, we noticed that cookies, third party, and track are used often. Also, financial information such as credit card, account details, transfer and transact is frequently discussed. Security of such data is very important and we have noticed that secure is highly used in shopping domain. Words such as market, promotion, offer, and ad are also used in the e-commerce policies commonly. Also, the option for unsubscribing is used most often in the e-commerce domain (23.16%). The other important interesting observation is that SSL, safeguard, socket layer and safe harbour are used frequently in shopping related policies.

**Kids:**

We were also interested to investigate the alignment of children online service privacy policies with regulations. According to COPPA, information regarding third party operators such as the advertising network or the social network that collect children’s information must be mentioned in the policy [58]. However, we noticed that third party is only used in 12.61% children-related privacy policies which is low in comparison to other domains. This suggests that children online service providers do not comply with the recommended practices as much as expected.

We extracted several lines from one of the children’s sites regarding their data collection practices [59]. This service provider is transparent about the data they collect from children. However, the type and sensitivity of collected data is concerning. The company also indicates that children’s personal information may be shared with third parties:

“We may collect and store information that you voluntarily supply to us while on our Site. This may include information that can be used to contact or identify you, such as your name, email address, or postal address. …

Location Information. We may collect location information (e.g., city and state and/or zip code) that you provide. Certain devices and browsers contain unique identifiers that can be used to identify the geographical location of the device. When you use such a device or browser to access our Site, your device and/or browser may automatically collect and/or transmit your location information. We also may collect GPS information from the device if the device transmits it …

We also collect and store information through a variety of automatic technologies, such as browser and flash cookies, pixels and web beacons about your use of our Site, including device identifying information such as IP addresses, MAC address or other device-specific alphanumerical ID numbers, the make and model of your device, the wireless provider associated with your device, …

If you choose to connect to our Site through a social network such as Facebook, we may automatically collect your profile information. These things
allow us, among other things, to improve the delivery of our web pages to you and to measure traffic on the Site …

The information we collect may be collected directly by us, or it may be collected by a third-party website hosting provider, or another third-party service provider, on our behalf.”

This sample policy suggest that children data collection practices are not any different than adults. Furthermore, use of ambiguous words such as *may* open up opportunities for data misuse and sharing. Security section only covers about 5% in children policies. Kids policies do not make a notable use of words such as *security, SSL, safeguard*, and *socket layer* which support our previous observation about limited coverage of security section. Lastly, online service providers should restrict their users to 13-year-olds and above. This practice has not been effective and, therefore, PII data collection raises major concerns in particular due to the lack of a mechanism to identify and protect children’s data.

To sum up, we analyzed the most frequent words, ambiguous words and topic coverage of a large number of privacy policies to gain better insight into the content of privacy policies. An overall observation is that FIP has made the recommended practices somewhat consistent across the globe and there are certain frequent keywords and ambiguous words used in the policy. With the comprehensive analysis of privacy policies, we can build an ontology which can capture varied knowledge and vocabulary of privacy policies which is discussed in the next chapter. This suggests that the ontology can be used in various applications and regulatory regimes if the policies are in English language.
Chapter 3   Ontology

With ontology, knowledge can be represented by defining relationships and classification of concepts for a particular domain [60]. In general, domain ontology is built using the vocabulary and the relations among the concepts of any particular domain. For an ontology which captures the domain knowledge related to privacy policies, must represent the vocabulary and relationship among the concepts of data practices and queries related to users’ privacy concern. In other words, ontology must have two important aspects: taxonomy and relationships. Taxonomy is useful in clustering entities which have common ontological characteristics. Chapter 2 described the process of analyzing the keywords in policies which will form our base for the taxonomy. In this chapter, we will discuss the relationship between the taxonomy, privacy categories and the privacy related queries.

The process of building an ontology can vary depending on the purpose and usage of application. However, the core steps of building the ontology remain the same which is discussed further in this chapter. There are various languages which are used to build an ontology. Some of the languages are Web Ontology Language (OWL), Resource Description Framework (RDF), and DARPA Agent Markup Language + Ontology Inference Layer (DAML+OIL). OWL is the commonly used language and is developed by W3C (World Wide Web Consortium). OWL ontologies are further subdivided into three sub-languages: OWL-Lite, OWL-DL, and OWL-Full. OWL-Lite is the least expressive language while OWL-Full is the most expressive language. OWL-DL is moderately expressive and is based on Description Logic (DL). It is known for decidable and automated reasoning which can check for any inconsistencies in the ontology [63]. Due to these reasons, we used OWL-DL for building our ontology.

In this chapter, we will discuss the methodology to build an ontology (Section 3.1), followed by the experimental setting (Section 3.2). Later on, we will discuss our results in Section 3.3. Lastly, we will validate our ontology with the competency questions, structural metrics and OPP 115 dataset (Section 3.4) and discuss the results (Section 3.5). An overview of this chapter is given in Figure 3.1 below.

3.1 Methodology

3.1.1 Purpose of Building Ontology

As previously discussed, the users are not inclined to read lengthy policies. To solve this problem, semantic approach can be used which can guide users to read the relevant sections of the privacy policies as it relates to users' concerns. To this end, an ontology can be used to capture domain specific knowledge and analyze the privacy policies semantically to find answers to a user query. The next step in future can be to build a tool which can act as a recommender system which can suggest concerns related to the user’s primary concern and highlight the relevant sentences in the policy.
3.1.2 Proposed Approach

We aimed to build an ontology to semi-automate the semantic analysis of privacy policies. Ontology is useful in capturing domain specific conceptual knowledge. It captures the vocabulary found in the privacy policies and map relationships between terms and concepts. These terms are mapped with privacy categories and competency questions. The competency questions, the users’ concerns written in plain English language, are further linked with SPARQL (SPARQL Protocol and RDF Query Language) queries in the ontology. With the mapping of the terms with the privacy categories and competency questions, the ontology acts as a knowledge base which can parse the privacy policies to find the relevant sentences related to the competency question and recommend those sentences only to the users as the answer to their queries.

3.1.3 Scope of Ontology

The proposed ontology is focused on the domain of privacy policies. The best way to determine the scope of the ontology is to determine the list of questions which the ontology is intended to answer. These set of questions, also known as competency questions, help in identifying the depth of the knowledgebase. By determining the competency questions, the required information can be included in the ontology in the form of terms or vocabulary. In short, the level of the granularity of the ontology is related to the competency questions and the terms identified for the ontology which will be discussed in the following sections.
3.1.4 Source of Ontology Development

As discussed in Chapter 2, privacy policies have inconsistent format and terminology. Therefore, the ontology must include diverse terms and concepts of privacy policies so that it is a comprehensive knowledge-base. For this reason, we have performed analysis of privacy policies to learn about the important keywords and content of policies. The findings of this study are discussed in detail in Chapter 2. These keywords were then used as the vocabulary to build the ontology. The other core components of our ontology are the privacy categories and competency questions. The common data practices used in the privacy policies were categorized into privacy categories as discussed in Chapter 2. The privacy categories used in the ontology are: Data Collection, Data Sharing, Choice, Access, Data Retention, Data Security, Policy Change, Do not Track, and Purpose. The ontology should satisfy the competency questions (Section 3.3.3) to represent the user’s privacy related queries. The competency questions were selected based on the 17 concerns that the users expressed about their online privacy [62].
This work is an extension of the ontology built by Audich et. al [62]. The authors built a small ontology with 35 vocabulary (keywords) by extracting keywords using supervised and unsupervised algorithms. The ontology was basic and covered only three competency questions.

3.1.5 Building Ontology-Steps and Process

We can build an ontology by following the steps mentioned in Chapter 1. The first step in building the ontology is to identify the terms or keywords. Most of the analyzed terms became the “ground level” components of ontology which are also called *individuals*. The terms with the common concept were grouped into classes, and a class hierarchy was formed. The class hierarchy can be built using top-down approach, bottom up approach and the combination of both the approaches. We used the combination of both approaches and organized the classes in the form of a taxonomy. Later, relationships between classes were built which helped in describing the semantics of the domain. Furthermore, attributes and instances were added to provide more context in the ontology. The terms were further mapped to the competency questions and data practices to get an in-depth insight of the domain knowledge. We have built the ontology in iterations focusing on the terms related to one query at a time. This helps in maintaining a large ontology and future expansion can be performed easily. To build an ontology, we used Protégé which is a free and open source ontology editor to build an OWL-DL ontology [64]. The common steps in building the ontology is demonstrated in Figure 3.2.

![Diagram of steps in building ontology](image)

**Figure 3.2 Steps in building ontology**

3.1.6 Competency Questions

Competency Questions (CQs) are a set of questions which are written in natural language format and the ontology should be able to answer the question based on the set of axioms
or mappings defined in the ontology. CQ play an important role in the development of ontology as they represent the ontology requirements as well as they are helpful in evaluating the ontology. The competency question in this work is written in plain English and is further translated in SPARQL.

First, we built the ontology based on some simple CQs which covered the basic concerns of users. Later, we combined two or more conceptual axioms to form some complex CQs which covered multiple concerns. The CQs are listed below (Table 3.1):

<table>
<thead>
<tr>
<th>Competency Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1: Which PII are covered by this policy?</td>
</tr>
<tr>
<td>Q2: Does this website track me?</td>
</tr>
<tr>
<td>Q3: Does this website use cookies?</td>
</tr>
<tr>
<td>Q4: Does this policy have an opt in/opt out policy?</td>
</tr>
<tr>
<td>Q5: Does this policy specify any data retention practices?</td>
</tr>
<tr>
<td>Q6: Will the user be notified of any changes to this policy?</td>
</tr>
<tr>
<td>Q7: Does this policy discuss about data security practices?</td>
</tr>
<tr>
<td>Q8: What choices are available in the policy?</td>
</tr>
<tr>
<td>Q9: Can I have access to my data?</td>
</tr>
<tr>
<td>Q10: Which PII are being shared under this policy?</td>
</tr>
<tr>
<td>Q11: Does this website share my personal information with third parties?</td>
</tr>
<tr>
<td>Q12: Does this policy discuss about tracking cookies?</td>
</tr>
</tbody>
</table>

### 3.2 Building an Ontology

This section presents the ontology which was built iteratively using the Ontology 101 methodology. Ontology 101 provides a step by step approach for building ontologies iteratively in Protégé [24]. It is a frame-based system in which objects are described along with the properties. The top structure of the ontology is explained in Section 3.3.1. The relationships of ontology are discussed in Section 3.3.2 and Section 3.3.3 includes answers to the competency questions.

#### 3.2.1 Structure of Ontology

The ontology is divided into seven classes namely Country, Keyword, Legal Acts, Organization, Privacy Category, Privacy Policy, and Question. Owl:Thing is the root class in the hierarchy of the ontology under which all the classes are defined. An example of a class is given in Table 3.2. Figure 3.3 illustrates the structural view of the ontology. The classes represent different concepts related to privacy policies and are discussed below:
1. Keyword class stores the terms which are mostly used in privacy policies. The keyword class is further divided into nine sub classes to describe the types of keywords (see Figure 3.3 and Figure 3.4). Furthermore, the sub classes are divided until the concept cannot be further divided. For example, the sub class informationType is further sub divided into contactInfo, deviceInfor, otherPI, userInfo which are the types of information. Under each sub-class, a set of individuals are listed which are the “ground level” components. Figure 3.4 illustrates an example of individuals for classes under information_type. Table 2 shows an example of a class along with the individuals.

2. Privacy Category has nine individuals: dataCollection, dataRetention, dataSecurity, dataSharing, dontTrack, policyChange, Purpose, userAccess, and userChoice which represents different types of privacy categories.

3. Question class contains all the questions as individuals. There are twelve different types of questions which represent some privacy concerns. Additional information such as concern, competency question, and SPARQL query is stored using data property which is also discussed in the next section.

4. The class Privacy Policy has two sub classes: Cookie Policy and Policy document which store the processed privacy policies. The policies specific to cookies will be stored as individual under the Cookie Policy sub class. Similarly, privacy policies will be individuals of Policy Document class.

5. The class Organization has two sub classes: Government and Private to store the respective types of organizations. For example, Federal Trade Commission is an instance of Government sub class.

6. Legal Acts class stores different types of acts as instances such as Children’s Online Privacy Protection (COPPA).

7. Country class captures different types of countries such as Canada and USA.

<table>
<thead>
<tr>
<th>Table 3.2 Example of a class in the ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
</tr>
<tr>
<td>Description</td>
</tr>
<tr>
<td>SubClass of</td>
</tr>
<tr>
<td>Instances</td>
</tr>
</tbody>
</table>
Figure 3.3 Classes in Ontology

Figure 3.4 Example of individuals for a class in ontology
3.2.2 Relations in Ontology

The ontology consists of is-a relations with the hierarchy of the OWL-DL ontology. Furthermore, additional relations were built to capture complex logic. This was performed by using object properties and data properties. Object property helps in describing the relationship between classes based on their instances to map complex logic. Table 3.3 shows the relationship between classes using the domain and range of the object properties. Similarly, data property helps in relating instances to a literal data. For example, retrieval date, sensitivity, URL, SPARQL, number of lines helped in capturing more information (Refer to Figure 3.5). The combination of object properties and data properties helped in improving the effectiveness of the ontology. This further helped in inferring useful information such as which keywords were related to a privacy category and its relevant query. Since the ontology captures the necessary information including the SPARQL questions, this can help in reducing time for further development and processing of policies.

Table 3.3 Object properties used in ontology along with the domain, range and characteristics

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Domain</th>
<th>Range</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>appliesTo</td>
<td>legalActs</td>
<td>Country</td>
<td>InverseOf: enacted</td>
</tr>
<tr>
<td>enacted</td>
<td>country</td>
<td>LegalActs</td>
<td>inverseOf: appliesTo</td>
</tr>
<tr>
<td>based</td>
<td>country</td>
<td>Organization</td>
<td>InverseOf: operatesIn</td>
</tr>
<tr>
<td>operatesIn</td>
<td>organization</td>
<td>Country</td>
<td>InverseOf: based</td>
</tr>
<tr>
<td>similarTo</td>
<td>keyword</td>
<td>Keyword</td>
<td>Transitive, Symmetric</td>
</tr>
<tr>
<td>relatedTo</td>
<td>keyword, Question</td>
<td>Privacy Category</td>
<td>Symmetric</td>
</tr>
<tr>
<td>Collects</td>
<td>informationType</td>
<td>dataCollection</td>
<td></td>
</tr>
<tr>
<td>has Access</td>
<td>accessRights</td>
<td>userAccess</td>
<td></td>
</tr>
<tr>
<td>hasChoice</td>
<td>choiceType</td>
<td>userChoice</td>
<td></td>
</tr>
<tr>
<td>notifies</td>
<td>notificationType</td>
<td>policyChange</td>
<td></td>
</tr>
<tr>
<td>haspurpose</td>
<td>purposeType</td>
<td>purpose</td>
<td></td>
</tr>
<tr>
<td>retains</td>
<td>retentionActions</td>
<td>dataRetention</td>
<td></td>
</tr>
<tr>
<td>secures</td>
<td>securityType</td>
<td>dataSecurity</td>
<td></td>
</tr>
<tr>
<td>shares</td>
<td>informationType</td>
<td>dataSharing</td>
<td></td>
</tr>
<tr>
<td>tracks</td>
<td>trackOptions</td>
<td>donotTrack</td>
<td></td>
</tr>
</tbody>
</table>
3.2.3 Competency Questions

The results from the Competency Questions can be used to validate the correctness of the ontology and is helpful in knowing the requirements of the ontology. Some of the competency questions along with the answers, privacy category and SPARQL query is given below and the rest of the competency questions are in Appendix B.

Q1: Which PII are covered by this policy?

**Query:** What are the terms related to personal information?

**Answer:** address, email, phone number, postal code, name, date of birth, username, age, password, credit card, bank account number, social security number, health card number, driver license number, IP address, location, personally identifiable information

**Category:** Data Collection
SPARQL:

Q3: Does this website use cookies?

**Query:** What are the terms related to cookies?

**Answer:** cookie, browser, gif, pixel tag, web beacon, disable

**Category:** Data Collection, Do not track

SPARQL:

Q11: Does this website share my personal information with third parties?

**Query:** What are the terms related to personal information and third party sharing?

**Answer:** address, email, phone number, name, date of birth, IP address, credit card, username, password, age, credit card, bank account number, social security number, health card number, driver license number, IP address, location, postal code, personally identifiable information, share, third party, advertiser, service provider, marketing partners, subsidiaries

**Category:** Data Sharing
3.3 Validation

Validating a domain ontology can be difficult in nature. There are different approaches that an ontology can be validated. These are briefly discussed in Chapter 1. We have performed validation of ontology in four ways: using CQs (Section 3.3.3), metric based validation, data driven validation using OPP 115 dataset and user evaluation. User evaluation was performed in this research and results are presented in detail in Chapter 4.

Validation using OntoMetrics

Evaluating ontology using metric based method can help in determining the quantitative quality of the ontology. This can help in gathering the numerical data related to the knowledge presented in the ontology. We used a web tool, OntoMetrics, which displays statistics about an ontology [61]. It is a metric based automated ontology evaluation for empirical evaluation. This can be useful for ontology users before they consider an ontology as a source of information. Furthermore, it can be used by ontology developers to evaluate the work of building ontology. Metric calculation is used to access the ontology characteristics. In OntoMetrics, the metrics are divided into several groups which are as follows:

1. Base metrics
2. Schema metrics
3. Knowledgebase metrics

Base metrics: Base metrics represent the quantity of ontology elements such as the number of axioms, classes, properties and more. Table 3.4 shows the list of base metrics which represent the number of elements of ontology. Axiom is a main component of an ontology which represent basic statements of an ontology. The axioms which impact the logical meaning of an ontology is referred to as logical axiom count. Class count helps in representing the concepts in an ontology. In other words, this metric helps in quantifying the number of classes in an ontology. Object property is useful in linking individuals of a
class to the individuals of other classes. Similarly, data property helps in linking individuals to the data values (literals). Individual count represents the number of instances in a class which represent the objects of a domain. A set of annotations which associates information in an ontology is described using the annotation property count. Description Logic (DL) is a formal knowledge representation language which describes the relevant concepts of a domain. In short, base metrics describe the count of basic properties of an ontology.

![Table 3.4 Base metrics](image)

**Table 3.4 Base metrics**

<table>
<thead>
<tr>
<th>Property</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axiom</td>
<td>525</td>
</tr>
<tr>
<td>Logical axiom count</td>
<td>275</td>
</tr>
<tr>
<td>Declaration axioms count</td>
<td>157</td>
</tr>
<tr>
<td>Class count</td>
<td>27</td>
</tr>
<tr>
<td>Object property count</td>
<td>15</td>
</tr>
<tr>
<td>Data property count</td>
<td>14</td>
</tr>
<tr>
<td>Individual count</td>
<td>96</td>
</tr>
<tr>
<td>Annotation Property count</td>
<td>8</td>
</tr>
<tr>
<td>DL expressivity</td>
<td>ALHIF+(D)</td>
</tr>
</tbody>
</table>

**Schema metrics:** Schema metrics are the metrics related to the design of the ontology. It is difficult to know if the ontology is designed correctly. However, with the help of these metrics, we can understand some important features such as the richness, depth, width and inheritance of an ontology (Table 3.5):

1. **Attribute Richness:** It is the number of attributes for all classes divided by the number of classes. Generally, more number of attributes means more knowledge is represented in the ontology.
2. **Inheritance Richness:** It is the average number of subclasses per class. It is an indication of well the knowledge is divided into categories and sub-categories in the ontology. Low Inheritance Richness means that the ontology is vertical in nature and covers a specific domain in detail. However, high Inheritance Richness means that the ontology is horizontal in nature and covers a wide range of knowledge.
3. **Relationship Richness:** It is the number of non-inheritance relationship divided by the total number of relationships in a schema. An ontology that has diverse relationships covers more knowledge as compared to the ontology that only has inheritance relationship. Relationship Richness represents the diversity of relations in an ontology.
4. **Axiom-Class ratio:** It is the average amount of axioms per class.
5. Class-Relation Ratio: It is the ratio of number of classes and the number of relations in an ontology.

Table 3.5 Schema metrics

<table>
<thead>
<tr>
<th>Property</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute Richness</td>
<td>0.518519</td>
</tr>
<tr>
<td>Inheritance Richness</td>
<td>0.740741</td>
</tr>
<tr>
<td>Relationship richness</td>
<td>0.428571</td>
</tr>
<tr>
<td>Axiom/Class Ratio</td>
<td>19.444444</td>
</tr>
<tr>
<td>Class/Relation Ratio</td>
<td>0.771429</td>
</tr>
</tbody>
</table>

By analyzing the results presented in Table 3.5, we can infer the features of the domain ontology. Axiom/Class ratio is 19.44 which shows that the number of axioms in the ontology is greater than the number of classes. Similarly, Class/Relation ratio is 0.77 which means that the number of relations is greater than the number of classes in the ontology.

Knowledgebase metrics: It is an important measure of ontology quality as it represents the amount of real-world knowledge in an ontology which is based on the design and how the data is placed in the ontology. The knowledgebase metrics assess the structure of the ontology as well as the instances that are populated in the ontology. The results are presented in Table 3.6.

1. Average Population: It is the ratio of number of instances of knowledgebase and the number of classes defined in the ontology schema. It compares the number of instances with the number of classes. The result represents the knowledge sufficiency using the extraction of instances to represent the knowledgebase.
2. Class Richness: It is the ratio of the number of classes with instances and the total number of classes. It gives a general idea of how well the knowledgebase utilizes the knowledge in the ontology. Low Class Richness means that the knowledgebase does not have the data that exemplifies the knowledge existing in the schema. High Class Richness means the data represents enough knowledge in the knowledgebase.

Table 3.6 Knowledgebase metrics

<table>
<thead>
<tr>
<th>Property</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Population</td>
<td>3.555556</td>
</tr>
<tr>
<td>Class Richness</td>
<td>0.814815</td>
</tr>
</tbody>
</table>
Validation using 2000 policies dataset

We used over 2000 policies to validate the ontology. We conducted two experiments [62] for validation which are as follows:

1. Policy Coverage: calculates the number of sentences that the reader has to read to understand the risk associated with his concerns.
2. Completeness: calculates per policy keywords that exist in the ontology but not in the policy

Experiment 1 (Policy Coverage): In this experiment, the coverage of sentences in the policy which addresses the query, or the privacy concern is calculated. The average of total sentences in 2000 policies was calculated and is represented as Total Sentences (TS). The number of sentences that included the keywords returned by the SPARQL query from the ontology was calculated and is denoted by Selected Sentences (SS). The policy coverage is calculated as follows:

\[
\text{Coverage} = \frac{SS}{TS} \times 100
\]

The results for all the queries are presented in Table 3.7. This experiment demonstrated that the user has to read some selected sentences ranging from 1 sentence to 29 sentences depending on the privacy concern. Query 1, 10 and 11 has the highest coverage as compared to other queries. This is because a greater number of keywords are associated to these queries. The rest of the queries has lower number of keywords which is discussed in Section 3.2.3.

<table>
<thead>
<tr>
<th>Query</th>
<th>Total Sentences (TS)</th>
<th>Selected Sentences (SS)</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>99.14</td>
<td>22.39</td>
<td>22.58%</td>
</tr>
<tr>
<td>2</td>
<td>99.14</td>
<td>1.05</td>
<td>1.05%</td>
</tr>
<tr>
<td>3</td>
<td>99.14</td>
<td>5.92</td>
<td>5.97%</td>
</tr>
<tr>
<td>4</td>
<td>99.14</td>
<td>2.57</td>
<td>2.59%</td>
</tr>
<tr>
<td>5</td>
<td>99.14</td>
<td>2.19</td>
<td>2.20%</td>
</tr>
<tr>
<td>6</td>
<td>99.14</td>
<td>3.58</td>
<td>3.61%</td>
</tr>
<tr>
<td>7</td>
<td>99.14</td>
<td>6.28</td>
<td>6.33%</td>
</tr>
<tr>
<td>8</td>
<td>99.14</td>
<td>3.88</td>
<td>3.91%</td>
</tr>
<tr>
<td>9</td>
<td>99.14</td>
<td>11.27</td>
<td>11.36%</td>
</tr>
<tr>
<td>10</td>
<td>99.14</td>
<td>25.99</td>
<td>26.21%</td>
</tr>
<tr>
<td>11</td>
<td>99.14</td>
<td>29.77</td>
<td>30.02%</td>
</tr>
</tbody>
</table>
Experiment 2 (Uniqueness): In this experiment, the number of keywords returned by the SPARQL query that did not exist in the privacy policies is calculated. Since the ontology is built using keywords from 2,500+ policies around the world, this experiment is useful in knowing how universal the ontology is. The uniqueness of the ontology is calculated as follows:

\[
Unique = \frac{NF}{KF + NF} \times 100
\]

Where KF is the average number of keywords found in the policies and NF is the average number of keywords not found in the policies. The results are presented in Table 3.8.

<table>
<thead>
<tr>
<th>Query</th>
<th>Keyword Found</th>
<th>Not Found</th>
<th>Unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.43</td>
<td>9.06</td>
<td>67.16%</td>
</tr>
<tr>
<td>2</td>
<td>0.48</td>
<td>1.10</td>
<td>69.62%</td>
</tr>
<tr>
<td>3</td>
<td>1.42</td>
<td>3.34</td>
<td>70.16%</td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
<td>2.96</td>
<td>74.74%</td>
</tr>
<tr>
<td>5</td>
<td>1.13</td>
<td>3.63</td>
<td>76.26%</td>
</tr>
<tr>
<td>6</td>
<td>1.04</td>
<td>0.54</td>
<td>34.17%</td>
</tr>
<tr>
<td>7</td>
<td>2.44</td>
<td>5.50</td>
<td>69.26%</td>
</tr>
<tr>
<td>8</td>
<td>1.61</td>
<td>2.35</td>
<td>59.34%</td>
</tr>
<tr>
<td>9</td>
<td>2.75</td>
<td>2.00</td>
<td>42.10%</td>
</tr>
<tr>
<td>10</td>
<td>5.35</td>
<td>10.53</td>
<td>66.30%</td>
</tr>
<tr>
<td>11</td>
<td>6.59</td>
<td>13.26</td>
<td>66.80%</td>
</tr>
<tr>
<td>12</td>
<td>1.90</td>
<td>4.45</td>
<td>70.07%</td>
</tr>
</tbody>
</table>

3.4 Discussion

Since ontologies can be complex in nature, various methods of evaluation and validation are required to evaluate different aspects of ontology such as the number of terms, depth, and density of taxonomy and relationships. By using a metric based automated ontology evaluation, we can access the quality of the ontology in terms of the components and structure of the ontology. Schema metrics helped in evaluating the design of the ontology and knowledge representation. Knowledgebase metrics was useful in checking whether the ontology was rich and had accurate representation of real-world entities and relations as it helped in evaluating the placement of instance data in the ontology. Overall, the results of both the schema metrics and knowledgebase metrics showed that the ontology captured the required knowledge related to the data practices.
As discussed earlier, the major concern is that users usually do not read long policies and reducing the text can help in cut down the time required to read policies and, in turn, can encourage users to read policies. With Experiment 1, we have calculated the number of sentences a user has to read for a given privacy concern. This experiment has showed that the user has to read on average about 1 to 30 sentences in a policy relating to the privacy concern. Furthermore, it is noted that the keywords related to personal information (PII like email, address) are used more often in the policies. This is confirmed by analyzing the results of Q1, Q10 and Q11 which had the highest policy coverage with the coverage of 22.58%, 26.21% and 30.02% respectively. On the contrary, the lowest policy coverage was found for Q2, Q4 and Q5 with the coverage of 1.05%, 2.59% and 2.20% respectively which also indicates that the keywords related to the respective practices are used less often. Overall, the findings show the significant reduction of sentences the user has to read for a query.

By further analysis and comparing the privacy queries, it is noted that the highest number of keywords returned by the ontology for any query is for Q1, Q10 and Q11. Q1 and Q10 are general queries related to PII and sharing of PII respectively. Q11 is a specific query related to third party sharing of personal information. Since the ontology returns highest number of keywords for these queries, the coverage is also the highest for these queries. Furthermore, it is noted that if we remove common words such as share then the sentences coverage decreases by 3%. This means that including generic keywords can increase the number of selected sentences for a query. Therefore, there should be a fine balance of keywords which is more directed towards the privacy query.

While building the ontology, we used the keywords from more than 2600 policies to make the ontology diverse and not restricted to a particular application domain or regime. We wanted to evaluate the ontology based on how unique the keywords are in the ontology. By analyzing Experiment 2, it is observed that depending on the privacy query, the keywords were unique ranging from 34% to 76%. By calculating the mean for all the queries, it is noted that the uniqueness is of 63.83%. This shows that the ontology captured diverse keywords.
Figure 3.6 Sentences selected for Q4: Does this policy have an opt in/opt out policy?

Figure 3.7 Sentences selected for Q7: Does this policy discuss about data security practices?
For further analysis of the sentence coverage in the policy, we plotted graphs to visualize the results. Figures 3.6, 3.7 and 3.8 show the policy coverage for Queries Q4, Q7 and Q11. The rest of the graphs are given in Appendix C. In Figure 3.6, it is noted that Survey Monkey policy has the highest Selected Sentences for Q4. The total number of sentences in Survey Monkey policy is 447 sentences and the Selected Sentences are 38 sentences. This signifies that for Q4 the user has to read only 38 sentences instead of reading the entire policy. Similarly, for Q7, PayPal had the highest Selected Sentences (Figure 3.7). The user has to read only 101 sentences as opposed to reading 2048 sentences in the policy. For Q11, Next Privacy policy had the highest Selected Sentences i.e. 321 sentences (Figure 3.8). It is noted that there is a reduction of 1403 sentences as the total number of sentences were 1724 in the policy.

In conclusion, with a greater number of sentences in the policy, the user has to read a lot to understand the privacy concerns. The keywords returned from the ontology can be helpful in highlighting sentences in the policy which can reduce the reading time of users. This is discussed in detail in Chapter 4 where a user study session was conducted to further evaluate the ontology.
Chapter 4   Evaluation of Ontology using User Study

As we have discussed in the previous chapter, evaluation of ontology can be performed in different ways. Ontology can also be evaluated by the users. This can be done by conducting user studies. There are many approaches in which the user study can be conducted which is discussed in Chapter 1. We used the task-based approach. The task-based approach helps in measuring how the ontology is useful in improving the results of a particular task. This can be performed by engaging users to access and evaluate the ontology based on some tasks.

In general, user study means systematic examination of the characteristics and behavior of the users of a system or service. It helps in understanding the users’ behaviors and needs. One of the major reasons for conducting user study is to measure the effectiveness of the ontology. The end goal is that the ontology is useful to the users and does the required work.

User study is useful to collect data from a sample of individuals to make inference about a wider population. It is one of the low-cost methods to generate a large amount of data in a short period of time [75]. User study is directly linked with the effectiveness of the system (or ontology in our case) as it focuses on user needs and demands. In this Chapter, we will present the details of our task-based approach for conducting user study. Section 4.1 gives an overview of task-based approach. Section 4.2 presents the methodology. Experimental setting is discussed in Section 4.3. The metrics, results and discussion are explained in Section 4.4, Section 4.5, and Section 4.6 respectively.

4.1 Task Based User Study

In task-based user study, users or individuals try to complete some real sets of tasks. The participants are observed while they are performing the tasks and the users evaluate qualitatively as well as quantitively their experience using some metrics. This can help in evaluating users experience during the tasks. There are certain guidelines in formulating a task [74]: 1) the tasks in the user study should be realistic and actionable in nature; 2) they should be easy to perform and should be clearly stated; and 3) the task should represent or cater to the overall goal or objective. Task-based user study can be of two types: Cognitive task-based, and Hierarchical task-based [73]. In the cognitive task based, the tasks require decision making and problem-solving skills. However, hierarchical task based is a decomposition of high-level task into simpler sub tasks.

A significant body of work has been done in the field of user study. Sushmita et al. [65] conducted user study that evaluated the effectiveness of an aggregated search interface for non-navigational search tasks. The participants were given situational background scenario of the task. After the end of the task, a questionnaire was given to capture their perceptions of the tasks. In another user study, the participants were asked to perform some simple tasks like writing an email, filling a form so that they can compare devices (laptop and tablet) to evaluate usability and user experience [66]. The task-oriented metrics such as effectiveness, efficiency, and ease of use were calculated. Furthermore,
device-oriented metrics such as usability and user experience were also calculated. Similarly, McDonald et al. [67] evaluated different formats of privacy policies using user study session. The participants were asked to compare six different policies and answer a set of questions. Based on their answers, analysis was done to calculate accuracy, speed results and psychological acceptability results (trust, and enjoyment). Hasse et al. used a task-based approach to evaluate the quality of the ontology. The authors measured the efficiency by how the ontology allows users to obtain relevant instances as answers for their search [79].

4.2 Methodology

4.2.1 Motivation and Objective

For a successful user study, the objectives or the goals should be clearly stated. The tasks should cater to the overall goal and should help in making useful inferences related to the objective. The tasks would be the basic foundation of the user study with which an overall objective can be achieved.

The primary goal of conducting user study sessions in this research was to evaluate and validate the effectiveness of ontology presented in Chapter 3 to find relevant content to a query in privacy policies. We used task-based user study approach to understand users’ perception and collect their feedback on the usability of ontologies. By using the task-based approach, users can perform tasks, based on which we can evaluate whether the ontology is effective in finding sentences in the policies related to the query. This approach is compared with users manually finding the information related to a query in the policy. In other words, we are interested in comparing the effectiveness of manually finding the relevant sentences related to a privacy query verses a semi-automated method using ontologies. This process is formulated into some simple tasks and the users can rate their experience while performing the tasks. This can help in identifying which approach works the best and is more user friendly. Through user study session, we were interested in examining the following:

1. Experience of users reading and understanding an entire policy.
2. Steps the user takes to find information related to a query in the policy and their experience in finding information. The possible steps can be thinking and searching of certain keywords which helps in finding relevant sentences for a particular privacy concern in the policy.
3. Experience reading the relevant sentences regarding a privacy concern which is identified using the semantic approach of ontology.

4.2.2 Proposed Approach

In the user study, participants are invited to perform some tasks in which each task caters to a specific objective discussed earlier (Section 4.2.1). These tasks are helpful in examining and comparing the experience of users finding relevant content in the policy
manually as compared to semi-automated semantic approach. This helps in identifying which task takes less time and is user friendly. A short questionnaire is provided to the users at the end of each task to capture users’ views related to the task and to understand their experience performing that task. At the end, a set of metrics is used to measure users’ experience.

4.3 Experimental Setting

4.3.1 Planning and Formulating Tasks

Planning the task is one of the most important aspects in user study as the task directly impacts the success of the study. The tasks act as the foundation of the user study and further analysis can be performed to evaluate the system under investigation as a whole. A task should reflect the research goal effectively and accurately [68]. Narrowing down the research goal into a set of tasks can be difficult. Furthermore, quantifying the users’ experience related to the tasks is also very crucial.

There are several factors that are important to take into consideration when formulating the tasks. The task must be simple and clearly written along with a task scenario. Task scenario helps put the task into context and motivates the user to participate in the study. Clear instructions must be given to the user. The tasks should be actionable and realistic [68]. Based on the above-mentioned objectives, we formulated three tasks which are as follows:

**Task 1**: A privacy policy is given to the user to read. This task is a warm up task in which the users can get familiar with privacy policy and its data practices by reading the policy.

**Task 2**: A privacy concern or a question is given to the user along with a privacy policy. The user has to highlight the sentences in the policy which the user feels answers the concern. It is suggested to the user to think of certain keywords related to the concern and search them in the policy to find relevant sentences.

**Task 3**: The user will be given the ontology highlighted sentences in the policy which covers the answer to the same privacy concern as mentioned above. The user has to read the highlighted sentences.

4.3.2 Selecting Policies and Queries

We were interested in using multiple privacy policies to collect diverse data from our users. We selected the policies based on some attributes: length of the policy, most visited sites and low-ranking policy. We considered 2000 policies data set to select privacy policies for the user study. The longest policy in the dataset was honda.com with 371 sentences. We selected Google.com as it is the most visited site according to Alexa.com. Electronic Frontier Foundation (EFF) rates the policies every year based on some factors such as transparency, informing users about government data requests, providing meaningful notice and had rated Amazon among the low-ranking policy.
Therefore, for the three tasks we finalized three different policies: Google, Amazon and Honda for our user study experiments.

With diverse policies, we want to have multiple queries or privacy concerns for the policy. For each policy, we selected a specific query or a privacy concern. This helped in collecting diverse data for the user study. The queries selected for the policy are listed in the Table 4.1.

<table>
<thead>
<tr>
<th>Policy name</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>Does this policy discuss tracking cookies?</td>
</tr>
<tr>
<td>Honda</td>
<td>Are users given a choice on opting-in/opting-out?</td>
</tr>
<tr>
<td>Google</td>
<td>What are the policies related to retention practices?</td>
</tr>
</tbody>
</table>

The queries are selected in such a way that it covers different data practices and the complexity of the queries vary. We started with a simple query to make the participants comfortable with the task. For the first policy Amazon, we wanted to select a simple query “Does this policy discuss tracking cookies?”. Since most of the policies have a separate section related to cookies, queries related to cookies can be much easier to find. Furthermore, we believe most users are familiar with keywords related to cookies and tracking and therefore, we expect users to find the sentences easier. The second and third queries can be difficult to find for the users as the keywords related to the queries may not be familiar to an average user.

### 4.3.3 Questionnaire

Developing a good questionnaire is crucial in collecting users’ experience and views related to the tasks in the user study. The questions must be simple, clear and must be focused towards the task or goal. The questions must be straightforward and written in simple language so that it is easy to understand. For our user study, we developed different questionnaires which included a pre survey questionnaire and questionnaire after the end of three tasks. The questionnaires are provided in the Appendix D.

Before the start of the user study, a pre survey questionnaire was given to the user which asks about basic information like education, computer literacy, internet literacy. It also included some questions such as concern regarding privacy and data practices, how often users read policies and whether they have made attempts to read an entire policy. The questionnaire included a combination of multiple-choice questions, polar questions, open ended questions and rating questions. These questions can help in knowing the user’s background and privacy preferences.

After each task, a short questionnaire was given to the user asking about their experience and views related to the task. The questionnaire consisted of open-ended questions, polar questions, rating and Likert scale questions. These questions can help in capturing the user’s views and their experience in the user study.
4.3.4 User Study Session Setup

According to Nielson Norman Group, testing five users is typically enough to identify a design’s most important usability problems [69]. Turner et al. have also shown that five participants can help in finding 80% of the usability problems [71]. Virzi made three claims regarding sample size for usability studies [72]: (1) Observing four or five participants allows practitioners to discover 80% of a product’s usability problems, (2) observing additional participants reveals fewer and fewer new usability problems, and (3) observers detect the more severe usability problems with the first few participants. There is no certain number of participants for finding all usability problems.

For our user study, we selected five participants based on their education, English proficiency and technical/non-technical knowledge. After selecting these participants, an invitation letter was sent giving details about the user study. It is important to note that this study was reviewed by the University of Guelph Research Ethics Board for compliance with federal guidelines for research involving human participants and the project ID is REB 16-12-245 titled “Building an Ontology for Privacy Policies 16MY022”.

4.3.5 Setting up User Study

A number of factors are considered before setting up a user study session. Some of them are selecting a moderating technique, deciding the metrics, equipment required in user study session which is discussed in this section.

Moderating the user study is important to know about the users’ insight towards the task. Some of the common moderating techniques are Concurrent Think Aloud (CTA), Retrospective Think Aloud (RTA), Concurrent Probing (CP), Retrospective Probing (RP). CTA is used to understand the thoughts of the participants by having them think aloud while they work on a task. The goal is to encourage participants to keep a running stream of consciousness as they work. In RTA, the participant is asked to retrace their steps after the session is completed. Generally, the participants watch a video replay of their actions performed in the task. CP is used to understand the thoughts of participants by asking them follow-up questions as they do something interesting or say something unique while working on the task. In RP, questions are asked about the participant’s thoughts and actions after the end of task. For our study, we used the combination of Concurrent Think Aloud and Retrospective Probing.

For conducting user study session, we used a laptop to perform all the tasks. To measure the time on task, we used a stopwatch. We kept the user study consistent with all the participants.

The process of user study

For each participant, the process of conducting the user session is described in this section. Each participant was welcomed and was asked to sign the consent form and fill the pre-survey questionnaire. The facilitator explained think aloud moderating technique
and explained the study session. The facilitator asked if the participant had any questions. The participant read the first task and its instructions and began to work while thinking aloud. The facilitator made notes on the side if there was any peculiar observation. After the end of the task, a short questionnaire was given asking about their experience with the task. The facilitator asked some questions regarding the task based on the observation made earlier. The facilitator also asked about the participant’s overall view or comments on the task. This process was repeated with task 2 and task 3. Then, the facilitator thanked the participant for their participation.

4.4 User Study Metrics

Metrics are a standardized method with which user’s experience can be measured. Based on the metrics, meaningful inferences can be made from the user study session and users’ experience. Quantifying the user’s experience by using metrics can provide insights towards the tasks and can help evaluate the ontology and evaluate whether the users prefer the highlighted text related to a privacy concern. Following are the metrics that used for this user study:

1. **Time on task**: The time taken to complete the task by the user is measured. It is helpful when we want to know how quickly users can perform tasks in the user study. Use of stopwatch will be required. The timer will be turned on at the start of each task and turned off after each task. Time will be recorded after each task and then average will be calculated for each task. Analysis can be performed based on how much each individual user will take to complete the task and overall average time for all the users for each task.

2. **Efficiency**: Efficiency is the ratio of the task completion rate to the mean time per task. Task completion rate is calculated by dividing the number of assigned participants who completed tasks successfully by the total number of assigned participants. This will help in knowing how efficient the task is. The mean can be calculated for each task.

3. **Ease of Completion**: It is used to measure task-performance satisfaction. We use the Single Ease Question (SEQ) “Overall, how difficult or easy did you find this task?” The question is asked after the completion of every task and participants have to rate the difficulty on a 7-step Likert scale from “very difficult” to “very easy”. Mean scores can be calculated to compare the difficulty of the tasks.

4. **Likert scale score**: It is used to assess participant’s satisfaction after completing each with a 7-point Likert scales (1=strongly disagree, 7=strongly agree). The score is calculated by taking the average of responses given by all the participants. If a question is skipped, then the remaining questions are considered for the average score. Mean score can be calculated to make inferences.
For task 1, we only considered one metric i.e. the time on task to measure how much time it takes to read the entire policy. For task 2 and task 3, we used all the above metrics to make comparisons for both the tasks and among all participants.

4.5 Results

In the following sections, we present the results of the user study. The first section will briefly depict the pre-survey questionnaire results and the later sections will present the result of metrics.

Pre-survey questionnaire

The results from the pre-survey questionnaire are given in this section which captures education, background and privacy concerns of the users. Five users participated in the user study in which 2 were undergraduate students and 3 graduate students (Masters: 1 and PhD:2). The participants were from various educational fields, namely, business, computer science, food science, and accounting. We further made the conscious effort to select participants from multiple background based on English proficiency, and technical/non-technical knowledge.

In the pre-survey questionnaire, the users reported their computer and internet literacy level which is presented in Figure 4.1. The graph clearly illustrates that the majority of the users rated their computer literacy and internet literacy as very good or good. Furthermore, we asked the users regarding their concerns about their privacy and data practices in the privacy policy. Figure 4.2 illustrates that all the users were concerned about their privacy.

![Figure 4.1 Computer literacy and Internet literacy of the users](image-url)
Furthermore, the users were asked if they have ever read an entire privacy policy. The results were quite surprising. Only 1 user reported that they had read an entire policy. The others stated reasons for not reading the entire policy. Three users stated that the privacy policies are too long. The other reasons mentioned were that privacy policies are hard to understand, time consuming, users do not have enough time to spend reading the entire policy, thereby, they save time by not reading them.

**Time to read the privacy policy**

We were interested in measuring the time to read privacy policies. Therefore, we formulated task 1 for users to read an entire policy and the results are given below. For Amazon policy, the 5 users took 11min 7 sec, 12 min 42 sec, 12 min 29 sec, 12 min 10 sec, 16 minutes with the average of 12 min 90 sec. Since the Honda privacy policy is too long, it took a lot of time as compared to other policies. The range of time for this policy was between 18 min 30 sec and 27 min with the mean of 22 min 60 sec. It took an average of 11 min 78 sec to read the Google privacy policy with the individual time ranging from 10 min 30 sec to 14 min 54 sec.

**Time on task**

We were also interested in comparing the time elapsed for completing task 2 and task 3 for the three policies. Figure 4.3, 4.4 and 4.5 illustrates the task completion time for Amazon, Honda and Google respectively. Comparing the results for each task for the three policies shows that task 3 took less time. For task 2, the users had to think about certain keywords relevant to the privacy concern which was time consuming as well as
overwhelming for some of the users. As a result, the completion time for task 2 was very high.

**Figure 4.3 Time taken by the users to complete task 2 and task 3 for Amazon policy**

**Figure 4.4 Time taken by the users to complete task 2 and task 3 for Honda policy**
Figure 4.5 Time taken by the users to complete task 2 and task 3 for Google policy

By comparing the mean completion time for each task shown in Table 4.2, it is clear that task 3 took the least amount of time to complete. Therefore, it will take the least amount of time if the users are given sentences highlighted according to the privacy concerns as compared to the whole policy document. This can help in encouraging users to read privacy policies.

Table 4.2 Mean time per task

<table>
<thead>
<tr>
<th>Task</th>
<th>Amazon</th>
<th>Honda</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>12.90</td>
<td>22.6</td>
<td>11.78</td>
</tr>
<tr>
<td>Task 2</td>
<td>7.36</td>
<td>3.80</td>
<td>2.90</td>
</tr>
<tr>
<td>Task 3</td>
<td>1.11</td>
<td>0.87</td>
<td>1.07</td>
</tr>
</tbody>
</table>

Efficiency

The efficiency of each task for the three policies is illustrated in Figure 4.6. Comparing the three tasks shows that task 3 is the most efficient for all the three policies because it takes the least amount of effort and time to complete.
Ease of use

Ease of Use was measured using the Single Ease Question (SEQ): “Overall, how difficult or easy did you find this task?”. The answers were rated on a Likert scale form 1 (very easy) to 7 (very difficult). Table 4.3 summarizes the results using the mean values instead of median because the number of participants was low.

<table>
<thead>
<tr>
<th>Task</th>
<th>Amazon</th>
<th>Honda</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 2</td>
<td>4.8</td>
<td>5.4</td>
<td>4.4</td>
</tr>
<tr>
<td>Task 3</td>
<td>1.4</td>
<td>1.2</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Results show that the users felt that task 3 is easier than task 2. The users were also asked reasons for their rating in the questionnaire. Users felt task 2 was difficult because of multiple reasons. They felt that the process was time consuming as the users had to identify the relevant keywords and concepts and then read the content. In fact, participants had to understand the legal terminology related to the privacy query which can be difficult for a non-technical person with no background related to law. Furthermore, analyzing too much information in a short period of time can be overwhelming. The users stated that task 3 was easy because the process was simplified by an ontology which saved significant time.
Likert scale score

Likert scale, also referred as satisfaction scale, is used by participants to rate the degree to which they agree or disagree with a statement. We used a 7-point scale to rate various statements or characteristics. We calculated the mean value to represent the scales for all the users. The statements are given below:

1. Meet your expectations
2. Effort to complete the task
3. Time to find the information

![Meet Expectations](Figure 4.7 Likert scale score for task 2 and task 3 based on expectation)
Results from Figure 4.7, 4.8 and 4.9 shows that the task 3 outperformed for all three attributes in comparison with task 2. The users felt that the highlighted text from task 3 met their expectations. Furthermore, the users experienced that task 3 took less effort to complete and less time to find relevant content.

4.6 Discussion

After analyzing the above results, it is clear that task 3 outperformed task 2 in term of task completion time, time to find relevant content, effort to complete the task, ease of use,
efficiency, and met the user’s expectations. At the end of the task, we further had a short talk with the users asking about their overall experience. There were some interesting observations made when the users were asked about their experience in performing task 1, task 2 and task 3. We also noted some interesting observations during the study session which is discussed in this section.

Task 1:

By analyzing the number of sentences in the privacy policies, it is clear that reading Honda policy would take the maximum time as it had the greatest number of sentences. The users reported that they would never be interested in reading such a long policy. The reading time was longer as the policy was lengthy. Similarly, the participants found task 1 to be time consuming for the other two policies: Amazon and Google. However, this observation does not necessarily mean that the length of policy should be reduced by removing the important information about data practices. The general perception of the users for task 1 is that it is quite time consuming to read an entire policy as the policy is too lengthy. Furthermore, the users also reported that they were unfamiliar with some of the terms used in the policy. e.g. safe harbor.

Task 2:

Query: Does this policy discuss tracking cookies?

The users had to find relevant keywords related to the query in Amazon privacy policy. This was an easy question because most of the policies explicitly discuss cookies and tracking in their policies. They have a separate section related to cookies in the policy document. Therefore, it was easier to identify sentences for cookies for task 2. Even if someone had no or little idea about cookies, they could search or skim through the policy to find the word “cookie” and find the relevant sentences.

Query: Can I opt in/ opt out of this policy?

The users were asked to find keywords related to this query in Honda privacy policy. It is interesting to note that some of the users were not aware of opt in and opt out data practices which was explained to them before the start of the task that helped them perform the task. Some of the users were able to find the relevant information, however, they missed some of the sentences as they were not able to identify the related keywords to opt in and opt out. Also, the relevant sentences were very less and were scattered throughout the policy. Finding sentences regarding opt in and opt out in the policy was similar to finding a needle in the haystack.

Query: Does this policy have data retention practices?

The participants used the Google privacy policy to identify the relevant sentences related to the query. Some users reported that it was difficult to highlight sentences related to the
query. This is because the sentences related to data retention were a little indirect and were not explicitly mentioned. Furthermore, the users had a hard time to think about keywords related to data retention. As a result, the users missed some sentences which covered the data practices related to data retention.

Overall, the users missed highlighting some of the sentences in the policies in task 2. Some of the reasons were scattered relevant sentences in the policy and difficulty in thinking of relevant keywords. Prior basic knowledge (or strong legal literary) is required to find the relevant information related to the data practices. Furthermore, it can be overwhelming for the user to find relevant content as the users still have to read the entire policy and understand it.

Task 3:

For all the three policies: Amazon, Google, and Honda, the users stated that task 3 was the easiest. The users commented that task 3 saves time by highlighting only the relevant information. By analyzing the total number of sentences in the policies, it is noted that Amazon, Honda, and Google had 152 sentences, 371 sentences, and 122 sentences respectively. With the help of ontology, the number of sentences for the respective query were reduced to 10 sentences (Amazon), 14 sentences (Honda), 9 sentences (Google). The users felt that task 3 is more user friendly as compared to the other tasks. It highlights the important information related to the privacy concern and makes it easy for the user to read less sentences. The users stated least effort is required and meets the user expectations.

Overall, the users stated that they preferred task 3 as compared to task 2 which is also proved by the results of Single Ease Question. Since in task 3 the users were assisted to find relevant content, least amount of effort was required and was efficient. This was achieved with the help of ontology which captured the domain knowledge that helped in directing the relevant content to the users. On the contrary, in task 2, the users had to think about relevant keyword and search in the privacy policy. This made it difficult for the users. The users felt task 2 was overwhelming. Also, it required some prior knowledge of data practices in the policies and some users were not familiar with the legal terms related to the privacy concern. To summarize, we evaluated the ontology by conducting user study and it is concluded that if users are given the relevant keywords or summarized text then the users are more likely to read privacy policies as compared to reading an entire policy. This will help them enhance their privacy literacy and be informed about the data practices in the policies.
Chapter 5  Conclusion

Privacy policies play an important role in informing users about how their data is collected, shared and managed. However, the policies are long and difficult to read for most of the users. It requires reading skill considerably higher than an average literacy level. Therefore, the users are not inclined towards reading privacy policies. Due to these reasons, users are not aware of the data practices in the policy. Previous attempts have been made to simplify policies to help the users in reading policies.

This thesis presents a semantic approach to assist users to find relevant sentences which are related to the users’ concern or query. This is achieved by using domain ontology. To build an ontology related to privacy policies, we firstly examined the content of privacy policies by performing some experiments. An ontology is built using selected keywords to identify the relevant sections of policies that user should read regarding the privacy concern. We further evaluated and validated the ontology using different approaches: using competency questions, metric based evaluation, 2000 policies dataset, and user evaluation.

Results of content analysis of privacy policies suggested that there are some common keywords which are used in the policies consistently. Furthermore, we also analyzed the coverage of policies to check which data practice sections are given more attention in policies. It was identified that, consistently and for all types of policies, collection has the highest coverage followed by purpose and sharing. The important sections that are given less attention are security and choice. Furthermore, we analyzed the term frequency of keywords and seed keywords to form the vocabulary for the ontology.

The primary purpose is to build an ontology to reduce the amount of reading by directing the users to relevant sentences of the policies for a question. It is observed that with ontology, the relevant sections that the user has to read is reduced to 1.05% to 30.02% depending on the query and its associated keywords. With this, the reading time is also reduced. Therefore, this reduction in the number of relevant sentences encourages users to read and learn about data practices.

The ontology captured vocabulary according to the data practices that is used by different organizations and FIPs. Furthermore, the vocabulary is collected from different application domain and regimes which makes the ontology unique. This is further proved by the Experiment of Uniqueness. Evaluation of ontology was also performed using metrics which investigated the structure and design of the ontology. Furthermore, we conducted user study to evaluate the ontology. Task based approach was used and the users had to complete a set of tasks to evaluate whether the ontology helped give relevant answer for a query. The results show that the users prefer reading highlighted sentences for a concern instead of reading an entire policy as it took less time and effort.
5.1 Limitations & Future Work

This thesis presents the use of ontology to reduce the reading time of users for reading policies. However, there are some limitations which are as follows:

1. The current version of ontology does not include synonyms of keywords. The ontology can be extended by including synonyms to make the vocabulary rich.
2. The ontology only covers wide range of queries related to the data practices. More specific and precise queries can be added to the ontology.
3. The user study session only included five participants. A larger number of participants with different demographics can show a wider prospective of the effectiveness of ontology.

These limitations can be addressed in future work. Other future directions may include:

1. A browser extension or a tool can be built which can highlight the relevant sentences for the users’ concerns. It can also act as a recommender system which recommends other concerns the users might have.
2. The ambiguous words can also be highlighted in the policies to make the users aware of the ambiguous content. This can be done using contextual analysis which analyses the word co-occurrence. With this, the policy creators can reduce the amount of ambiguity in the content.
3. With the recent changes to new regulations like General Data Protection Regulation (GDPR), the content of privacy policies may have changed. The updated keywords or concepts can also be added to the ontology.
REFERENCES

47. Made “Made’s approach to privacy” https://www.made.com/made-privacy-policy
   (accessed February 10, 2018)
48. Information Commissioner’s Office “Subject Access Request”
   (accessed February 5, 2018)
   (accessed February 10, 2018)
    (accessed February 5, 2018)
51. Office of the Privacy Commissioner of Canada “Public Opinion Survey”
   (accessed February 5, 2018)
52. Office of the Privacy Commissioner of Canada “Consent”
   (accessed February 5, 2018)
   (accessed February 10, 2018)
54. Privacy in e-commerce “Privacy in e-commerce: a general perspective”
   https://privacypolicypinzai.blogspot.com/2015/03/privacy-policy-ecommerce.html
   (accessed February 5, 2018)
55. Thomson Reuters Practical Law “Data Protection in United States: overview”
   https://uk.practicallaw.thomsonreuters.com/6-502-0467?transitionType=Default&contextData=(sc.Default)&firstPage=true&bhcp=1
   (accessed February 5, 2018)
56. MyFitnessPal “Privacy” https://account.underarmour.com/privacy#information-we-collect
    (accessed September 2017)
    (accessed February 10, 2018)
58. Federal Trade Commission “Children’s Online Privacy Protection Rule”
   http://www.ftc.gov/ogc/coppa1.htm (last accessed December 26, 2017)


Appendices

APPENDIX A: Content Analysis of Privacy Policies

A1: Data protection laws in Canada, USA, UK, and Germany

Most countries have established regulations and data protection laws to govern and regulate privacy policies. However, the laws and regulations can vary between different countries. In the following few paragraphs, we provide an overview of the common privacy protection laws in Canada, USA, UK, and Germany. It is important to note that we limited our study to privacy policies in English.

In Canada, OPC enforces two federal laws: Privacy Act and Personal Information Protection and Electronic Documents Act (PIPEDA) (“Summary of privacy laws in Canada”, n.d.). Privacy Act deals with how the federal government handles personal information. On the other hand, PIPEDA helps govern personal information collected and used by businesses and private sector organizations (“Collecting Personal Information”, n.d.). Apart from the federal laws, some provinces of Canada have private sector privacy laws which are substantially similar to PIPEDA. For example, Ontario, Alberta, and British Columbia have Personal Information Protection Act (PIPA), which also governs the protection of personal information. Additionally, there are some domain specific laws which are regulated in some provinces of Canada. For example, Personal Health Information Protection Act (PHIPA) covers health related privacy laws in Ontario. Similarly, federal Bank Act covers provisions regulating the use and disclosure of personal financial information by the financial institutions (“Summary of privacy laws in Canada”, n.d.).

In USA, there is not a single federal law regulating the protection of personal information. USA has a patchwork of federal and state laws which can sometimes overlap or contradict with one another. In other words, the privacy laws in USA follow a sectoral approach in which the laws are specific to an industry sector and protect certain type of information (“US Privacy Law”, n.d.). In USA, FTC plays a major role in regulating business-related privacy laws (Earp et al., 2005). FTC principles are focused on prohibiting deceptive or illegal practices and work on five key principles: notice, access, choice, security, and enforcement (Earp et al., 2005; Westin, 1968). Due to the sectoral approach of privacy laws, there are some domain specific laws as well. For example, Health Insurance Portability and Accountability Act (HIPAA) deal with the protection of health-related information (“Summary of HIPPA security rules”, n.d.). Similarly, Children’s Online Privacy Protection Act (COPPA) helps govern the online collection of data from children under the age of 13 (“Children’s Online Privacy Protection Rule”, n.d.). Furthermore, there is a law to protect the Non-public Personal Information (NPI) of financial institutions like banks, security firms and insurance companies which is called Gramm-Leach-Bliley Act (GLBA) (Mamun, Hassan, & Maroney, 2005).
In UK, ICO is supportive in upholding rights for the interest of public (“Data Protection”, n.d.). Data Protection Act 1998 helps facilitate in governing transparency on how individuals’ data is protected. The act also prominently deals with the use of cookies. According to the act, the companies should explicitly explain why they are using cookies and should take informed consent from the users (“Data Protection”, n.d.). There isn’t any domain specific law in UK. However, ICO also enforces Freedom of Information Act 2000 (FOI) and Privacy and Electronic Communications Regulations (PECR). Under the FOI Act, the public can have access to information held by public authorities (“What is Freedom of Information Act”, n.d.). PECR Act gives people specific privacy rights in relation to electronic communications. Under this act, there are specific rules for cookies, marketing by electron means, security of public electronic communication services (“What are PECR”, n.d.).

In Germany, Federal Data Protection Act (Bundesdatenschutzgesetz in German) ensures that any personal information, even IP Address cannot be collected unless the user gives consent. According to the act, third party sharing is illegal, and data can only be collected directly from the users only with their consent.

Since there is disparity in national legislations, data transfer of personal data across countries could be difficult. To harmonize the national privacy laws and regulations, OECD was formed which works in developing guidelines and upholding human rights to build a national legislation (“OECD Guidelines”, n.d.). OECD works on various principles such as collection limitation, data quality, use limitation, purpose specification principle, openness, and accountability. These principles can be applied separately on national and international levels.

A2: Similarity analysis

Our observation from previous experiments suggested that there may be a large degree of overlap between the content of online privacy policies regardless of the regulation framework that they are created for or the application domain they are made for. We wanted to test this observation and, therefore, used a similarity metrics to measure overlaps between the words used in policies.

To calculate the degree of content overlap between policies, we first pre-processed the text and manually reviewed keywords to remove unrelated terminology. We then performed pair-wise comparison of word similarity between the following three groups of policies:

- C1: 2000 policy corpus and regulations
- C2: Canada, USA, and UK
- C3: 2000 policy corpus and application domains policies

Table 0.3 summarizes the results for experiments C1-C3.
As it can be seen in Table 2.3, there is a large degree of variation in keywords when we compare 2000 policy corpus with other smaller size policy corpora. This was expected since some of the policies in 2000 corpus are very long; they have 10000 words before pre-processing. This resulted in significant variation between keywords when comparing these policies even though we pre-processed all policies. When comparing overlaps between regulations, we noticed that USA and Canada policies are 41% similar. However, there was less overlap between European policies with USA and Canada.

We repeated the above experiments for seed keywords. As expected, Jaccard similarity measure confirmed that the extracted seed keywords are common legal terminologies in online privacy policies. Jaccard measure resulted in 90%-93% similarity between 2000 and different regions privacy policies, 90%-93% similarity between different regions privacy policies, and 94%-99% similarity between 2000 and application domain privacy policies. The largest dissimilarity score belonged to Canadian policies and largest similarity belonged to Kids and Health privacy policies (98% and 99% similarity to 2000 policies).
APPENDIX B: Additional Competency Questions used in the ontology

Q2: Does this website track me?

**Query:** What are the terms related to tracking?

**Answer:** IP address, track

**Category:** Do not track, Data Collection

**SPARQL:**

```sql
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX op: <http://www.semanticweb.org/jasmin/ontologies/2019/3/untitled-ontology-88#>

SELECT DISTINCT ?term
WHERE {
  (?term a op:trackOptions; op:tracks op:doNotTrack)
}
```

Q4: Does this policy have an opt in/opt out policy?

**Query:** What are the terms related to opt in and opt out?

**Answer:** opt in, opt out, subscribe, unsubscribe, consent

**Category:** User Choice

**SPARQL:**

```sql
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?term
WHERE {
  (?term a op:opt; op:hasChoice op:userChoice)
}
```

Q5: Does this policy specify any data retention practices?

**Query:** What are the terms related to data retention?

**Answer:** backup, database, delete, keep data, record, retain

**Category:** Data Retention
SPARQL:

Q6: Will the user be notified of any changes to this policy?

Query: What are the terms related to notifying changes?

Answer: notice, update

Category: Policy Change

SPARQL:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?term
WHERE {
  {?term a op:retention_actions; op:retains op:dataRetention} }
```

Q7: Does this policy discuss about data security practices?

Query: What are the terms related to security of data?

Answer: SSL, encrypt, socket layer, safeguard, compromise, secure, authorize, Secure Socket Layer, protect, fraud

Category: Policy Change

SPARQL:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?term
WHERE {
  {?term a op:notification_type; op:notifies op:policyChange} }
```

Q8: What choices are available in the policy?

Query: What are the terms related to choice?
**Answer:** choose, choice, option, wish, consent

**Category:** User Choice

**SPARQL:**

```sparql
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl1#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?term
WHERE {
  {?term a op:choice_type; op:hasChoice op:userChoice}
}
```

Q9: Can I have access to my data?

**Query:** What are the terms related to access?

**Answer:** delete, edit, preference, request, update, access

**Category:** User Access

**SPARQL:**

```sparql
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl1#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?term
WHERE {
  {?term a op:access_rights; op:hasAccess op:userAccess}
}
```

Q10: Which PII are being shared under this policy?

**Query:** What are the terms related to PII and sharing?

**Answer:** address, email, phone number, name, date of birth, IP address, credit card, username, password, age, credit card, bank account number, social security number, health card number, driver license number, IP address, location, postal code, personally identifiable information, share

**Category:** Data Collection, Data Sharing
Q12: Does this policy discuss about tracking cookies?

**Query**: What are the terms related to tracking and cookies?

**Answer**: IP address, track, Cookie, browser, gif, pixel tag, web beacon, disable

**Category**: Data Sharing, Data Collection

**SPARQL**:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?term
WHERE {
  {?term a op:contactInfo; op:shares op:dataSharing}
  UNION
  {?term a op:device_info; op:shares op:dataSharing}
  UNION
  {?term a op:otherPI; op:shares op:dataSharing}
  UNION
  {?term a op:userInfo; op:shares op:dataSharing}
}
```
APPENDIX C: Additional Results for Ontology Validation

Figure C.0.1 Sentences selected for Q1: Which PII are covered by this policy?

Figure C.0.2 Sentences selected for Q2: Does this website track me?
Figure C.0.3 Sentences selected for Q3: Does this website use cookies?

Figure C.0.4 Sentences selected for Q5: Does this policy specify any data retention practices?
Figure C.0.5 Sentences selected for Q6: Will the user be notified of any changes to this policy?

Figure C.0.6 Sentences selected for Q8: What choices are available in the policy?
Figure C.0.7 Sentences selected for Q9: Can I have access to my data?

Figure C.0.8 Sentences selected for Q10: Which PII are being shared under this policy?
Figure C.0.9 Sentences selected for Q12: Does this policy discuss about tracking cookies?
APPENDIX D: Survey Questions

D1 Pre- survey Questionnaire

Q1: Please indicate your education background

   a) No high school
   b) Some high school
   c) High school graduate
   d) Some college - no degree
   e) Associates/2 year degree
   f) Bachelors/4 year degree
   g) Graduate degree - Masters, PhD, professional, medicine, etc.

Q2: How would you rate your computer literacy (the ability to use the computer)? Please choose one of the following that best applies:

☐ Very Poor   ☐ Poor       ☐ Acceptable    ☐ Good       ☐ Very Good

Q3: How would you rate your Internet literacy (the ability to use the Internet)? Please choose one of the following that best applies:

☐ Very Poor   ☐ Poor       ☐ Acceptable    ☐ Good       ☐ Very Good

Q4: In general, how concerned are you about the protection of your privacy? Please circle one of the following scales:

Not concerned at all  1  2  3  4  5  6  7  Extremely concerned

Q5: How often do you read the privacy policies for Internet sites you visit? Please choose one of the following that best applies:

☐ Always      ☐ Often      ☐ Sometimes    ☐ Rarely     ☐ Never

Q6: Have you ever read an entire privacy policy?

☐ Yes           ☐ No

If not, what are the reasons for not reading the entire policy?
Q7: In general, how concerned are you about the data practices in privacy policies? Please circle one of the following scales:

Not concerned at all 1 2 3 4 5 6 7 Extremely concerned

**D2 Questionnaire after Task 1**

Q1: Which of the following categories best describes your experience reading the privacy policy? Please choose one option only.

☐ Very pleasant
☐ Somewhat pleasant
☐ Neither pleasant nor unpleasant
☐ Somewhat unpleasant
☐ Very unpleasant

Q2: Which of the following categories best describe your ease in understanding the data practices? Please choose one option only.

☐ Very easy
☐ Somewhat easy
☐ Neither easy nor difficult
☐ Somewhat difficult
☐ Very difficult

Q3: How would you rate the policy based on the following attributes? Please choose one of the options that best applies:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>(very short) 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>(very lengthy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of the policy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>(very easy)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>(difficult)</td>
</tr>
<tr>
<td></td>
<td>language</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time consuming</td>
<td>(less time)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>(more time)</td>
</tr>
</tbody>
</table>

**D3 Questionnaire after Task 2**

Q: What keyword/s are you specifically looking for?
Q1: Were you able to find answer to the query?

☐ Yes ☐ No

Q2: If not, what are the reasons for not finding the information regarding the query?
______________________________________________________________________

Q3: Was it easy to find the information? Please choose one of the following options:
☐ Very easy
☐ Somewhat easy
☐ Neither easy nor difficult
☐ Somewhat difficult
☐ Very difficult

Q4: How would you rate your experience of finding answer for the query based on the following aspects? Please circle one of the scales that best applies:

Meet your expectations

(No, it didn’t) 1 2 3 4 5  (Yes, I was expecting similar answer )

Effort to complete the task

(Less effort) 1 2 3 4 5  (More effort)

Time to find the information

(Less time) 1 2 3 4 5  (More time)

Q5: Overall, how would you rate this task?
very easy 1 2 3 4 5 6 7 very difficult

Q6: What is the reason that you gave the above rating?
______________________________________________________________________

D4 Questionnaire after Task 3

Q1: Did you find it easy to identify the information with the highlighted text? Please choose one of the following options:
☐ Very easy
☐ Somewhat easy
☐ Neither easy nor difficult
Q2: Do you feel that the highlighted text has covered the answer to the query?

☐ Yes  ☐ No

Q3: If not, what are the reasons or problems?
______________________________________________________________________

Q4: How would you rate your experience of having highlighted answer for the query based on the following aspects? Please circle one of the scales that best applies:

<table>
<thead>
<tr>
<th>Meet your expectations</th>
<th>(No, it didn’t) 1 2 3 4 5 (Yes, I was expecting similar answer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effort to complete the task</td>
<td>(Less effort) 1 2 3 4 5 (More effort)</td>
</tr>
<tr>
<td>Time to find the information</td>
<td>(Less time) 1 2 3 4 5 (More time)</td>
</tr>
</tbody>
</table>

Q5: Overall, how would you rate this task?

very easy 1 2 3 4 5 6 7 very difficult

Q: What is the reason that you gave the above rating?
______________________________________________________________________