Comparing and Analyzing Definitions in Multi-jurisdictions

Sepideh Ghanavati
Institute of Software Research, Carnegie Mellon University[1]
Luxembourg Institute of Science and Technology[2]
[1]Pittsburgh, Pennsylvania, US and [2]Luxembourg
sepideh.ghanavati@list.lu

Travis D. Breaux
Institute of Software Research
Carnegie Mellon University
Pittsburgh, Pennsylvania, US
breaux@cs.cmu.edu

***Abstract*— Regulatory definitions establish the scope and boundary for legal statements and provide software designers with means to assess the coverage of their designs under the law. However, the number of phrases that serve to define this boundary in a legal statement are usually large and often a simple legal statement contains or is affected by up to 10 definition-related phrases. In addition, software designers may need to design their software to operate under multiple jurisdictions, which may not use the same terminology to express conditions. Thus, it is necessary for designers to keep track of definitions in one or more regulations and to compare these definitions across jurisdictions. In this paper we report a study to develop a method to analyze and compare natural language definitions across legal texts and how to analyze the legal statements with respect to definitions. Our method helps reduce the number of comparison between definitions across multiple jurisdictions as well as allows software designers keep track of several inter-related definitions in a systematic way.**

***Index Terms*—Privacy, Legal Ontology, Definitions, Legal Requirements.**

I. Introduction

The trend towards service-oriented computing provides new opportunities for reuse, as software designers can leverage existing services to meet their own design needs. In heavily regulated domains, such as healthcare and finance, this introduces challenges of how to comply with laws across multiple jurisdictions. In healthcare, this is further complicated when people relocate and health data can be accessed in multiple locations: each governed by different laws. In the US, there are both Federal and state healthcare laws and some of these are interweaved inside other state codes and statutes. State laws range from public health monitoring of disease to medical record access and retention. Understanding how these laws define covered entities and information types can help software designers to determine the scope of the law and its applicability to the organizations, individuals, and supporting information technology (IT) systems.

While laws often list definitions in a separate section, comparing definitions across laws is still quite challenging. For example, consider the following definition from the Health Insurance Portability and Accountability Act (HIPAA) §160.103:

Health information means any information, including genetic information, whether oral or recorded in any form or medium, that:

(i) Is created or received by a health care provider, health plan, or public health authority, employer, life insurer, school or university, or health care clearinghouse; and

(ii) Relates to the past, present, or future physical or mental health or condition of an individual or the provision of health care to an individual, or the past, present or future payment for the provision of health care to the individual.

The subject of this definition is "health information," which could be reasonably compared to other definitions of health information to determine how multiple jurisdictions affect software design. However, the definition could be re-topicalized, which means we could change the subject and thus the emphasis to a different noun phrase without changing the meaning to yield two other definitions for: 1) health care providers who create health information; or, 2) an individual's past, present, or future payment for health care. Under these three definition topics, the designer might approach their coverage assessment by checking whether their design includes health information, or whether a use case includes primary actors working in a health care provider, or whether a use case concerns payments for health care. These three approaches on how definitions are topicalized direct the designer's attention to three different design perspectives: information types, an actor's institutional role, and business-related actions, all within the system's scope. Thus, a key challenge in comparing laws across jurisdictions is the problem of re-topicalization, or how to orient the constraints in a definition from different viewpoints in order to compare definitions from a shared viewpoint.

In this paper, we present our preliminary work on analyzing and comparing definitions. As a product of this process, we present an ontology for definitions that can be used by an analyst to classify definitions in ways that minimize the number of pairwise comparisons needed to find similarity and dissimilarity. To help perform the comparison and itemize the sentence topics into units, we extended the Frame-Based Requirements Analysis Method (FBRAM) [4]. Finally, we use a graphical notation to further aide in the comparison across legal texts, as

the number of possible comparable phrases increase with more complex definitions and multiple jurisdictions.

The remainder of the paper is organized as follows: Section II describes related work; Section III provides an overview of the research method; the research results are presented in Section IV, with the discussion in Section V. We summarize threats to validity in Section VI, and conclude with future work in Section VII.

II. Related Work

A. Frame-Based Requirements Analysis Method

The Frame-Based Requirements Analysis Method (FBRAM) [3][4] is a tool-supported manual annotation process. FBRAM uses concepts from an upper ontology [17], phrase heuristic and a context-free markup language to annotate legal statements and create legal requirements. The upper ontology has three types of concepts as statement-level concepts, phrase-level concepts and concept codes. Statement-level concepts such as obligation, permission and refrainment are used for categorizing every single legal statement. Phrase-level concepts such as act, subject, object and purpose are used to identify and annotate phrases in one single legal statement. Concept codes help mapping statements and phrases to these concepts. With the help of the context-free markup language, it is possible to codify an interpretation of regulations with upper ontology concepts and removing logical, attributive and referential ambiguities.

In our work, we extend FBRAM with the concepts from the definition ontology that we developed and use similar approach as FBRAM to codify the definition statements in regulations for further analysis.

B. Other Related Work

Comparing regulations in multi-jurisdictional environment and across several regulations have been the topic of many research in recent years. Ghanavati et al. [7] introduce a Goal-oriented based method for pairwise comparison of multiple regulations. In their method, they compare actors, preconditions, exceptions, modal verb, clauses and cross-references in legal statements and identify 6 cases of similarities, differences or conflicts. However, their method is manual and to do a proper comparison, they need to compare each legal statement with the other legal statements in the rest of the document. The authors expanded this work with the integration with Eunomos legal knowledge management framework [2], to semi-automatize part of the process [8].

Gordon et al. [9][10][11] developed a water marking framework to compare regulations across multiple jurisdictions in the US. In this framework, regulations are first translated into a canonical format using requirements specification language (RSL) [5] and then with the help of requirements watermarking approach and a set of metrics [5] for comparing two requirements, they analyze the differences and similarities between statements. To reconcile the differences and conflicts, the authors identified three heuristics (i.e. union, disjoint and minimum). In their work, the authors concentrate on comparison of legal statements without considering the definitions. In this work, we focus on comparing definitions.

Nómos 3 [15], an extension to *Nómos* [21] and *Nómos 2* [20], is a conceptual framework for representing legal norms and reasoning about the compliance with those norms. Norms in *Nómos 3* are defined as a 5 tuple of "type" of norms (such as Duty/rights), "holder" and "its counterpart", which are both legal norms, "antecedence" and "consequences" which are preconditions to make the norms applicable and post-conditions of the compliance. *Nómos 3* further uses the following concepts to establish compliance: Situation, which illustrates the state of the affair, for example "data concerning health information"; Goal which is a desired state of the affair, domain assumption which defines the scope of the norm and its relevant domains; Norms as type of right or duty and Roles as type of legal or social roles [14]. These concepts are used for analyzing and reasoning about compliance. Compliance is done by backward and forward inference reasoning and via traceability links between *Nómos 3* and *i** models. Up to now, extracting *Nómos 3* concepts and modeling them are done manually. Some of these concepts such as domain assumption and roles are usually find in the "Definition" sections of the regulations. Our method can be used to help *Nómos 3* in identifying the domain and role concepts in their models.

Massey et al. apply probabilistic topic modeling approach [1] to 2061 policy documents to identify the topics of the policy documents [16]. With the help of their approach the analyst can use a set of keywords to identify which of the policy documents contain the privacy protection and vulnerabilities. Their work is similar to us in the way that they try to topicalize policy documents. However, in our work, we do not only try to retopicalize the definitions based on the concepts but also we aim to compare the definitions from multiple jurisdictions with each other and to provide method on how to handle the definitions in a legal statement.

For identifying similarities in English text, there are databases such as Wordnet [23] and Framenet [6]. Wordnet and FrameNet are both large lexical databases of English language. Wordnet clusters nouns, verbs, adjective or adverbs into a set of cognitive synonyms, which are linked to each other by conceptual semantic and lexical relations. FrameNet includes annotated examples of how words are used in actual text. In our work, we can use either Wordnet or FrameNet to identify similar noun-phrases, verb-phrases or propositions to simplify the definition comparisons. However, since their corpora are not based on legal texts, we cannot use them to identify similarities between legal terms.

III. Definition Comparison Methodology

We now describe our approach to analyze and compare definitions across multiple jurisdictions. Our method consists of definition extraction and selection criteria, an ontology for categorizing selected definitions, a frame-based mark-up for annotating definitions, and finally a comparison of annotated units. In this paper, we only present preliminary results towards our goal of comparing and analyzing definitions across multiple jurisdictions and re-topicalization of the regulations. Figure 1

illustrates the steps for analyzing and comparing definitions across multiple regulations. In step A, we manually extract the definitions from the text of regulations. In step B, the category of each definition is identified manually based on the ontology we have developed. The definitions in each category are codified manually by a set of phrase-level concepts and they are compiled automatically to catch syntax and semantic errors such as introducing unknown concept codes or missing brackets in FBRAM tool in step C. In step D, we model definitions automatically in a graphical model and we compare the graphs with each other and identify the similarities and differences. In the following, we explain each of these steps in detail. Note that, step B and C can be done parallel to each other. By annotating the definitions in FBRAM, the categories from step B can be revisited and revised.

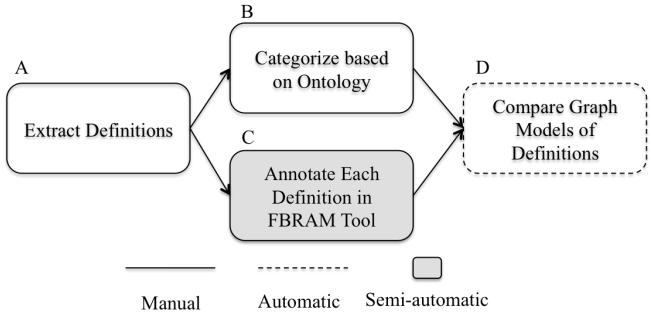

Fig. 1. Process for Definitions Comparison

We now discuss each step, separately.

A. Definition Extraction and Selection Criteria

In step A, we extract the definitions from selected regulations manually. In this paper, we focus on the federal and state healthcare regulations in US. Because definitions may not always appear in a separate section, we search for indicative keywords in the text as follows: "definition", "defined (as)", "mean(s)", "is synonymous (with)", "refers to" and "includes". We arrived at this list by analyzing HIPAA and the NY regulations, which already had definition sections.

B. Definition Categorization based on Definition Ontology

In step B, we apply the *definition ontology* to classify the definitions by the lead topic, which often may be an actor, an action, an information type, and so on. The ontology was first created using a grounded theory approach [13] and the ontology creation process defined by Rosch [22] (Rosch discusses on how to create an ontology based on the basic-level categories, which share the most common attributes in compare to their superordinate or subordinate categories). This step is important in the comparison of definitions as it helps reduce the number of $O(n^2)$ pairwise comparisons for n total definitions to a smaller number of $O(m^2)$ comparisons for m definitions in the most inclusive category. More distinct categories will reduce the number of comparisons in step D, at the cost of having more categories into which the definitions must be classified in this step. Thus, the method depends on striking the right balance, which we have only begun to explore, here.

The ontology consists of the following categories:

- *Action* means a statement that describes a noun phrase or a verb phrase related to performing an activity or an action. An example of an action is "disclosure" or "authenticate".
- *Actor* means a statement that defines person, individual, agent, organization or group of people. For example, "covered entity" or "family member" are actors.
- *Document* is a statement that describes the set of requirements, rules, conditions, agreements and plans. For example, "health plan" or "standard" are documents.
- *Facility* is a statement that defines a building or a physical place that used for provision of health care. An example of facility is a hospital.
- *Information* is a statement that contains any kind of information, data, records of information and notes containing information. "Medical information", "electronic health record" or "dataset" are types of information.
- *Technology* includes any statement that describes software, hardware, applications and technical methods. Technology can be for example, "password" or "technical safeguard".
- *Others* includes any definition that has the population less than 2%.

The ontology was constructed using grounded analysis by first analyzing the HIPAA definitions. During this initial classification, the main categories emerged as general topics under which the subject of the definition most easily fit (e.g., a "covered entity" is a kind of "actor"). Later, we identified the need to merge some categories or to introduce new categories. For instance, having two separate categories for personal and technical information is not useful as there are many types of information and some types, such as de-identified health information, can fall into multiple categories, which makes the classification in step B more arduous to repeat and risks incomplete classifications due to analyst fatigue. Thus, we decided to retain only the generic *information* category.

In addition, we encountered new terms, such as "password" and "malicious software" that did not fall into any previously identified categories. Thus, we created a new category for *technology*. Similarly, "resident," "patient service center" and "hospital" are not actors, in which case we introduced the category of *facility*. The definitions that are less than 2% of the total number of definitions and do not fit in any of the other categories are put in the *others* category. This category is intended to remain small, but may be a source of discovering new definition categories as new regulations are analyzed over time. For example, the term "direct treatment relationship" from HIPAA §164.501 states: "a treatment relationship between an individual and a health care provider that is not an indirect treatment relationship," is classified under others, because it does not fit under the other categories. Other subcategories under others include: abbreviations, terms, purposes, rules and penalty.

C. Frame-based Approach for Definitions

Recall, definitions may have more than one topic. In step B, we classify definitions by their primary topic based on the subject of the sentence or clause in which the definition is found.

To improve our ability to compare definitions under each category, we also applied and extended the FBRAM [3][4] approach. The FBRAM allows an analyst to annotate a regulatory document with phrase-level annotations. To that end, we defined a set of phrase-level legal concepts as part of a coding frame as follows:

- *Principal* (P) is the core concept in the left-hand side of a definition statement, which is restricted, by all other concepts. For example, "information" in the "health information" is a *principal*.
- *Definition* (D) is the clause in the right-hand side of a definition statement, which describes the meaning of the principal and its constraints.
- *Constraint* (C) is a word or phrase that puts restrictions on the principal.
- *Actor* (A) is the subject of the definition, or is any person, organization, agent, or group of people who is involved in or related to the principal.
- *Sub-type* (ST) is the sub-class of an actor or a constraint.
- *Cross-Reference* (CR) is a word or phrase that refers to other parts of the legal documents either internally or externally.
- *Connector* (CN) is a group of verb phrases or proposition phrases that connect the constraints, actors or cross-references to principal.

Figure 2 shows an excerpt from HIPAA §160.103, which we previously discussed in Section I. Markup is shown in **blue color** and the phrases are coded based on the phrase-level concepts discussed in this section. We use square [] bracket for *value blocks* to indicate the span of text that is annotated using our coding frame, above. The English conjunctions "and" and "or" are mapped to "&" and "|" to denote logical-and or logical-or, respectively. If a word or phrase is annotated as *Connector* (CN), the value blocks related to the connector are nested together in another value block to denote the relationship between both blocks in this binary relationship.

```
34          3 - [#C Health] [#P Information]- means [#D any information, [[#CN
35          including] [#C genetic] information, [[#CN whether] [[#C oral]| or
36          [#C recorded in any form or medium]]]], that:
37              (1) [[[#CN Is created]| or [#CN received by]] [[#A a health care
38              provider]|, [#A health plan] | [#A public health authority] |
39              [#A employer] | [#A life insurer] | [#A school or university] |
40              [#A health care clearinghouse]]] &
41              (2) [[#CN Relates to] [#C the past, present, or future physical
42              or mental health or condition of an individual] | [#C the
43              provision of health care to an individual] | [#C the past,
44              present, or future payment for the provision of health care to
45              an individual]].]
```

Fig. 2. An Example of Legal Definition Annotated with FBRAM

In Figure 2, the *principal* of the definition is "Information", which is annotated by #P, and which would indicate this definition should at least be classified under the information category in our definition ontology. Information is *constrained* by certain modifiers, including health, genetic, oral or recorded, which are annotated by #C and are nested by |. The *actors* related to this information *principal* are health care provider, employer, and public health authority, which are annotated by #A and are nested by |. Finally, *constraints* and *actors* are connected to the principal by *connectors* (a verb-phrase, gerund or proposition), which are annotated by #CN.

As mentioned earlier, the *principal* is the lead noun phrase or sentence subject and every other annotated value block extends from the principal. A natural first starting point in step D is to compare definitions with similar *principals*. The frame-based annotation, however, provides a richer vocabulary for performing comparisons by itemizing constraints and actors, their inter-relationships via connectors, the types and subtypes one might infer from these phrases, and additional context linked by legal cross-references [18]. In addition, since the value blocks are all linked to the principal and due to the usage of context-free markup language, ambiguities [4] in the definition statements can be removed.

D. Graphical Model for Comparison

In step D, we now compare the definitions using a tool for graphically modeling the annotated statements. The graphs consist of nodes and edges as follows: principals are the main nodes; and constraints, actors, subtypes and cross-references are linked to the principal node via edges. If some of the concepts are nested in connector with logical-and or logical-or, they are modeled in the graph using a "helper" node, which illustrates the type of nodes connected to it in the conjunction or disjunction. For example, if there are multiple actors grouped by a connector with logical-and, those actors will have a helper node called "actor", which is connected to the principal. The links between the nodes are attributed by the connector's annotated phrases.

After creating the graphs, we can analyze and compare the graphs. The first point of comparison is on principals. If the two principals are semantically similar, we start comparing constraints, actors, subtypes and cross-references. To do the comparison, we follow the steps below:

- **Step 1** – Connect the nodes that are exactly the same in both statements. For example, if the two statements are about "health information" and "de-identifiable health information", the node [#C Health] in both of the graphs shall be connected to each other.
- **Step 2** – Normalize connectors to simplify them. We can use Wordnet to identify the connectors that have semantic and lexical relations, group them together in one and replaced them all by one connector, which is semantically similar to them. For instance, the connectors [#CN pertaining] and [#CN regarding] can be replaced by [#CN about].
- **Step 3** – Connect the nodes that are semantically similar. For example, in the comparison between medical information and health information graphs, [#C Medical] and [#C Health] are semantically similar and can be connected to each other. This step is done manually and it is based on the judgment of the analyst. However, it can be possible to use Wordnet to assist comparison.
- **Step 4** – For the remaining nodes, we check if the following relationship exists between them which follows the same approach for comparing requirements based on phrase-level metrics [5]:

1) Subsumption: If actors, constraints or subtypes from one model subsume the actors, constraints or subtypes of the other model respectively. In this case, we can create subsumed links between the nodes of the similar concept.
2) A phrase-level concept subsumption to another concept: If an actor subsumes one or more constraints or a constraint subsumes one or more actors. In this case, we can create subsume links between the nodes of two different concepts.
3) No relation – The nodes are not similar or subsumed and they have nothing in common. In this case, if the designer needs to comply with both laws, the union of these nodes must be added to the analysis.

After comparison, the software designer can assess the design against the definitions and identify the scope of the legal statements for coverage. For example, if there is a use case for creating health information that applies to California, the designer can check if any of the legal statements apply to the use cases based on the definitions involved such as information or actors in both California and Federal laws and identify the relationship between these definitions and the use case.

IV. Analysis of the Result

In this section, we discuss our findings based on applying the comparison method discussed in Section III in a case study.

A. Definition Extraction and Selection Criteria

We selected healthcare regulations in the US for our case study to minimize the effect of dissimilarities among regulations from different countries in this initial study. In the US, the minimum standard for protecting the personal health information is the Health Insurance Portability and Accountability Act (HIPAA). Thus, we chose HIPAA as the base regulation for comparison. We selected state regulations based on two factors: 1) gross state product (GSP), which is the state-equivalent of gross domestic product; and 2) healthcare expenditures (HE). Table I shows the list of the top 10 states based on GSP and HE.

TABLE I. List of States based on GSP and Healthcare Expenditure

GSP	CA	TX	NY	FL	IL	PA	OH	NJ	NC	GA
HE	CA	NY	TX	FL	PA	IL	OH	MI	NJ	MA

As seen in Table I, the first seven states from left-to-right are ranked similar in GSP and healthcare expenditure. For our study, we focus on the first three states: California, New York and Texas. On January 1st 2015, California introduced the Confidential Health Information Act (SB 138) to close gaps in the HIPAA, which we added to our analysis. We extracted definitions from the following nine laws:

- **Federal Law**
  - HIPAA
- **CA (California)**
  - *Confidential Health Information Act (SB 138*
  - *Civil Sate Code*
  - *Welfare and Institution Code*
  - *Health and Safety Code*
- **NY (New York)**
  - *NY Code – Rules and Regulations Title 10 and*
  - *Public Health Law – Section 18*
- **TX (Texas)**
  - *Texas Medical Privacy Act (S.B.11)*
  - *Texas Health and Safety Code*

We next extracted definitions from each regulation using keyword search as discussed in Section III. This step was done manually as the effort was minimal. In total, we acquired 371 definitions. The number of definitions found per jurisdiction are shown in Table II.

TABLE II. Distribution of Definitions in Each of the Regulations

Regulation	**HIPAA**	**CA**	**NY**	**TX**
Total # of Definitions	113	94	80	84

B. Classification of Definitions and Information Category

Next, we classified the definitions using the categories in Section III.B; summary results appear in Table III.

TABLE III. Distribution of Definitions under Each Category

Definition Categories	**Total # of Definitions**
Action	**91 (24.5%)**
Actor	**121 (32.62%)**
Document	31 (8.35%)
Facility	14 (3.77%)
Information	**64 (17.25%)**
Technology	8 (2.16%)
Others	42 (11.32%)

The top three categories are actor, action and information. The *others* category contains all sub-categories with less than 2% of values. In classifying the definitions, we observed the following considerations:

- Elements such as digital signatures could fall under both *Information* and *Technology* categories. However, we chose *Technology* as their category and reserved the category *Information* for any type of information, record or data that needs to be protected by the law.
- Actions appear as nouns or verbs. Definitions of services are also categorized as actions, such as "radio therapy services", "child welfare services" or "sensitive services".

This categorization reduces the number of pairwise comparison, significantly. In our study, without categorization, we would have 371*370 pairwise comparisons of definitions, which is equal to 137,270 total comparisons. Based on our categorization shown in Table III, we reduced this number to (91*90 + 121*120 + 31*30 + 14*13 + 64*63 + 8*7 + 42*41) = 33,664, comparisons, which is 76% less than the original total.

Next, we annotated 64 *Information* definitions from the 9 mentioned regulations with FBRAM. From the 64 definitions in the *Information* category, we identified 12 unique principals. Eight principals appeared only once, the other four principals

appeared multiple times. The distribution of the principals are presented in Table IV:

TABLE IV. DISTRIBUTION OF PRINCIPALS IN INFORMATION CATEGORY

Principals	Total #
Information	27
Record	22
Data	5
Notes	2
Others	8
Total	**64**

As it is shown in Table IV, 76% of the principals identified in *Information* category are either "information" or "record." Further analysis of the left-hand-side (LHS) of definitions illustrates that there are 84 constraints attached to the principals on their LHS, while nine principals have no constraints attached to their LHS. Among these 84 constraints, the most common ones are "health" (16 times), "personal" (9 times) and "medical" (8 times). The constraints "patient," "protected" and "electronic" appeared 4 times each. These constraints are, without any exceptions, attached to either information or record. Furthermore, the term "medical" is used 7 out of 8 times in CA regulations whereas "health" was used once in the NY regulations and the rest of the time in HIPAA and TX regulations. This analysis shows the need to align constraints across jurisdictions, such as medical in CA and health in HIPAA, NY, and TX, to ensure that relevant definitions are compared.

Table V shows a tabular comparison between the three codified statements from HIPAA §160.103, NY §50-4.2 and CA §56.05.j, which are annotated in FBRAM. Each column corresponds to one definition, and each row appears in the order of the word or phrase in the original definition. The LHS constraints are shown in gray color and principals are shown in bold.

Analysis of the RHS of each of the definitions illustrates the similarities and differences. CA §56.05.j has one extra constraint [#C any individually identifiable] in compare with the other two statements. There are several connectors that are semantically similar to each other, as well. These connectors are: [#CN Including], [#CN Concerning], [#CN Whether], [#CN In], [#CN Is created or received by], [#CN In possession of or derived from], [#CN Relates to], ([#CN Which identifies] | [#CN could reasonably be used to identify]), and [#CN Regarding]. To simplify the analysis of the statements, we normalize the connectors. For example, instead of [#CN Including], [#CN Concerning], and [#CN Regarding] we can use [#CN Is about], because the principal of all of the three is information. We can also normalize [#CN Is created or received by] and [#CN In possession of or derived from] to [#CN In possession of] and [#CN Relates to] and [#CN Which identifies] to [#CN Relates to].

For the rest of the comparison, we follow Step 4 of Section III.D:

1) Subsumption – which is the case when one constraint (#C) or an actor (#A) can be subsumed from another constraint or actor, respectively. In Table V, column CA §56.05.j, the format of medical information is constrained by [#C electronic] or [#C physical form] while column HIPAA §160.103 has a constraint on the information of [#C oral] or [#C recorded in any form or medium]. Oral or recorded in any form can include electronic or physical form. Thus, we have a subsumption (>) relationship between the two constraints

TABLE V. COMPARISON BETWEEN DEFINITIONS IN HIPAA, NY AND CA LAWS

HIPAA §160.103	NY Public Health Law, §50-4.2	SB138 - CA Civil Code, §56.05.j
[#C Health]	[#C Personal]	[#C Medical]
	[#C Health-related]	
[#P Information]	**[#P Information]**	**[#P Information]**
		[#C any individually identifiable]
[[#CN including] [#C genetic]]	[[#CN concerning] [#C Health of a Person]]	
[#CN whether] [[#C oral] | [#C recorded in any form or medium]]		[#CN in][#C electronic] | [#C physical form]
[#CN Is created or received by]		[[#CN In Possession of] | [#CN Derived from]]
[#A health care provider] | [#A health plan] | [#A public health authority] | [#A employer] | [#A life insurer] | [#A school or university] | [#A health care clearinghouse]]]		[[#A provider of health care] | [#A health care service plan] | [#A pharmaceutical company] | [#A contractor]]
[[#CN Relates to]	[[[#CN which identifies] | [#CN could reasonably be used to identify]]	[#CN regarding]
[#C the past, present, or future physical or mental health or condition of an individual] | [#C the provision of health care to an individual] | [#C the past, present, or future payment for the provision of health care to an individual]].]	[#A a person]	[[#C a patient's medical history] | [#C treatment] | [#C mental and physical health condition].

2) Mix relationship – In Table V, HIPAA §160.103 and CA §56.05.j have one actor in common, [#A a health care provider] and [#A a provider of health care], while NY law has no such actor. The other actors of HIPAA §160.103 and CA §56.05.j are different. For example, [#A school or university] and [#A pharmaceutical company] are neither similar nor exist in a subsumption relationship.

3) An actor or a noun subsumes one or more constraints – In this case, an actor or a noun in one law can be expanded to a more detailed constraint in another law. In Table V, the information type has been defined in all the three laws. The NY §50-4.2 describes information that [#CN Is about] [#A a person]. This actor subsumes the constraints in the other two laws, which for HIPAA §160.103 are: [[#C the past, present, or future physical or mental health or condition of an individual] or [#C the provision of health care to an individual] or [#C the past, present, or future

payment for the provision of health care to an individual]], and, for CA §56.05.j, they are: [[#C a patient's medical history] or [#C treatment] or [#C mental or physical health condition]].

Further analysis between the constraints in HIPAA §160.103 and CA §6.05.j shows that constraints can fall in a "mix" group: [#C the past, present, or future physical or mental health or condition of an individual] subsumes [[#C a patient's medical history] and [#C mental and physical health condition]] and [#C the provision of health care to an individual] are subsumed by [#C treatment]. However, [#C the past, present, or future payment for the provision of health care to an individual] does not exist in CA §56.05.j.

C. Graphical Model for Comparison of 3+ Definitions

The tabular format in Table V provides a simple way to compare three similar definitions. However, as we compare more definitions, or as the comparisons are ordered linearly as they were in Table V, a graphical model provides more flexibility. In this step, we generate the graph automatically. Figure 3 shows the graph for HIPAA §160.103 and CA §56.05.j. The principal of the two graphs is [#P Information]. Both of the models have two constraints that are not nested, one set of nested actors and one set of nested constraints with two helper nodes called Actor and Constraint.

The comparison step is done manually. We start from the nodes that are directly connected to the principal node. Similar to Table V, we identified that "medical" and "health" are semantically similar and thus, the two nodes are linked to each other. Some nodes such as "health care provider" and "provider of health care" as well as "patient's medical history" and "past/present/future physical or mental health or condition of an individual" are connected to each other due to semantic similarities.

Beside this, from Figure 3, we also identify that there are three actors (i.e. health care provider, health plan and health care clearinghouse), which are defined in HIPAA as separate definitions and, thus have their own separate graphs. This is what we call internal cross-reference where a definition refers to another definition in the same document. With the graphical model, it is possible to create links between the graph of health information and the graphs of health care provider, health plan and health care clearinghouse. Figure 4 shows the link between definition for health care provider and definition for health information. With this connection, we end up having a hyper graph consist of the two sub-graphs (i.e. health information graph and health care provider graph) with the links between their common elements. The hyper graph includes the characteristics and constraints of both of the sub-graphs and the software designer must ensure to satisfy the requirements of all of the sub-graphs including in the hyper graph. Hyper graph can help software designers to maintain the relationships between definitions together. For example, if a software designer deals with a use case related to health information, with the help of hyper graph, the software designer can decide if the actor, health care provider, is relevant to the use case or not. One of the main benefits of the graphical model for comparison of the definitions is to help track the definitions in a legal statement and creating links between more than two definitions. This is an important challenge since legal statements usually refer to more than one definition. If the number rises to more than two internally cross-referenced definitions, it becomes difficult for analysts to keep all of the definitions in mind, to understand the scope of the legal statements with respect to the definitions correctly and to have the most coverage of the legal requirements. We analyzed 20 sections of HIPAA with 45 subparts and we identified that each legal statement contains at least four and maximum 14 linked phrases. Furthermore, as we see in Figure 4, a definition can refer to other definitions and, thus, this number can rises further.

.

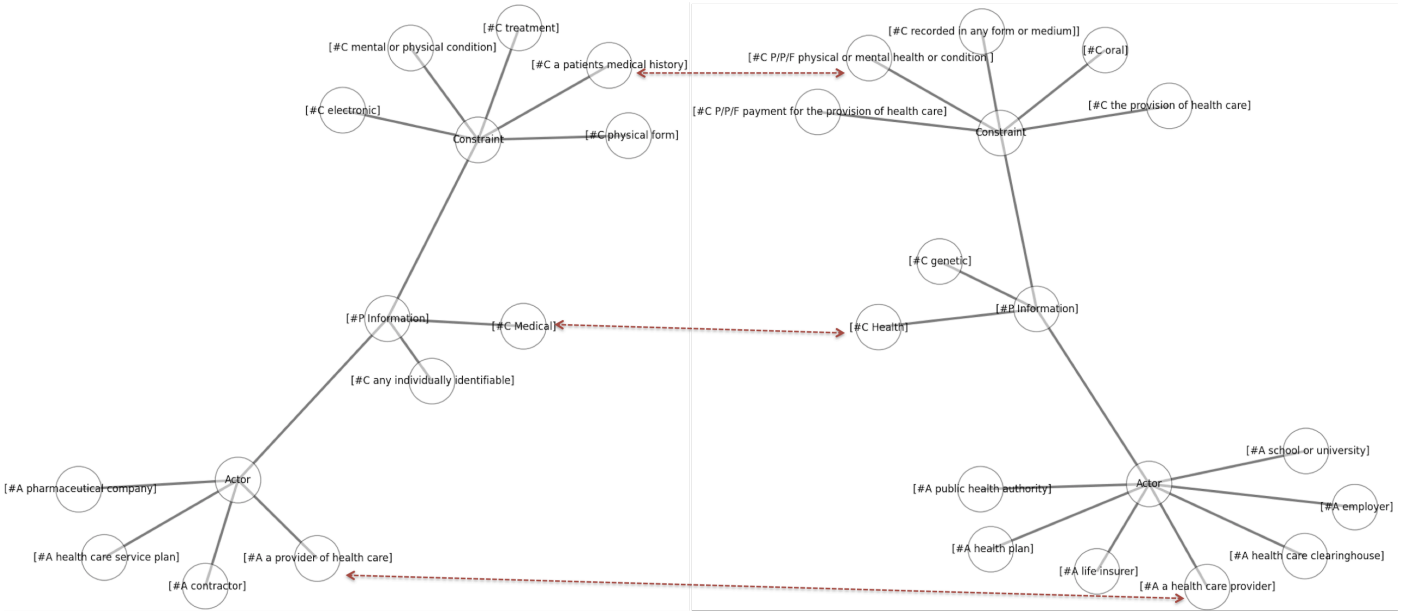

Fig. 3. Comparison between HIPAA and CA Law - Health Information vs. Medical Information

Fig. 4. Relationship between Health Information and Health care Provider

Consider §164.508 (b)(4) with prohibitions on the conditioning of authorizations:

A covered entity may not condition the provision to an individual of treatment, payment, enrollment in the health plan, or eligibility for benefits on the provision of an authorization, except: (i) A covered health care provider may condition the provision of research-related treatment on provision of an authorization for the use or disclosure of protected health information for such research under this section.

In this legal statement, we identified 10 definitions which are shown as underlined. Table VI presents the description of these 10 definitions. Few of these definitions have also cross-reference to other definitions in HIPAA, which are underlined in Table VI. We identified 10 new cross-referenced definitions in Table VI, which increases the number of definitions an analyst needs to consider while analyzing §164.508 (b)(4) to 20.

V. Discussion

As we note in Section IV, legal statements can become very complicated when we include the definitions. The graphical model can help improve our ability to compare definitions and create the links between similar constraints and actors on principals. However, in one statement, for example, we identified 20 definitions that an analyst needs to keep track of. Thus, software designers need better methods to trace and simplify legal statements and their integrated definitions.

TABLE VI. Definitions Mentioned in Article §160.308 (a) of HIPAA

Definitions	Description
Covered Entity	(1) A health plan. (2) A health care clearinghouse. (3) A health care provider who transmits any health information in electronic form in connection with a transaction covered by this subchapter.
Individual	Person who is subject of protected health information.
Treatment	Provision, [..] of health care and [..] by one or more health care providers, including [..] with a third party [..]
Payment	(1) The activities undertaken by:(i) [..] a health plan to obtain [..] for coverage and provision of benefits [..]; or(ii) A health care provider or [..] (2) The activities [..] relate to the individual to [..]
Health Plan	An individual or group plan that provides, or pays the cost of, medical.
Health care Provider	Provider of services [..], provider of medical or health services [..],any person or organization [..]
Use	Wrt individually identifiable health information, the sharing, employment, application, utilization, examination, or analysis of such information within an entity that maintains such information.
Disclosure	Release, transfer, provision of access to, or divulging in any manner of information outside the entity holding the information
Protected Health Information	Individually identifiable health information: (1) Except as provided in paragraph (2) of this definition, that is: (i) Transmitted by electronic media; (ii) Maintained in electronic media; or (iii) Transmitted or maintained in any other form or medium.
Research	Systematic investigation, including research development, [..]

In general, we include definitions to understand the scope of the legal statements and help ensure complete regulatory coverage [12] in our requirements. After the integration, we observed that we may be able to simplify definitions using the following heuristic: we remove constraints and actors from a definition to improve the comprehension while reaching the following states:

- No additional permission is given.
- No obligation has been excluded.
- No prohibition has been excluded.

To do the simplification, we have to consider all of these three cases. Table VII presents proposed rules for performing this simplification. In this table, O is for Obligation, P for Permission and ¬P for Prohibition, a and b are actors and p is an action described in a regulatory text. We can extend these rules to include information or document, as well. For simplicity in this paper, we only show the simplification for actors and actions.

Consider HIPAA §164.512 (k)(5)(ii) - Permitted uses:

A covered entity (CE) that is a correctional institution (CI) **may** use protected health information (PHI) of individuals who are inmates for any purpose for which such PHI may be disclosed.

This statement describes a legal permission and includes eight definitions. To simplify, we apply the rules in Table VII:

CE subsumes CI ($CI \subseteq CE$) and individual subsumes inmates ($Inmates \subseteq Individual$). Since the statement is a permission statement, we have to ensure that we give permission to only the groups that are allowed by this statement. In this case, only a subset of CE, which are also CI have $P_{CI}(Use)$ PHI of only a subset of individuals that are inmates ($P_{Inmates}(Use)$). Thus, rule 6 applies for both cases and, if we remove CE and individual, we still do not violate the legal statement and we can reduce the number of definitions to six.

TABLE VII. RULES FOR SIMPLIFYING LEGAL STATEMENTS

#	Type of Legal Statement	Action
1	$O_a(p)$ and $(a \subseteq b) \Rightarrow O_a(p)$	Remove b
2	$O_{(a \vee b)}(p)$	Keep a and b
3	$O_{(a \vee b)}(p)$ and $(a \subseteq b) \Rightarrow O_a(p)$	Remove b
4	$O_{(a \wedge b)}(p)$	Keep a and b
5	$O_{(a \wedge b)}(p)$ and $(a \subseteq b) \Rightarrow O_b(p)$	Remove a
6	$P_a(p)$ and $(a \subseteq b) \Rightarrow P_a(p)$	Remove b
7	$P_{(a \vee b)}(p)$	Keep a and b
8	$P_{(a \vee b)}(p)$ and $(a \subseteq b) \Rightarrow P_b(p)$	Remove a
9	$P_{(a \wedge b)}(p)$	Keep a and b
10	$P_{(a \wedge b)}(p)$ and $(a \subseteq b) \Rightarrow P_a(p)$	Remove b
11	$\neg P_a(p)$ and $(a \subseteq b) \Rightarrow \neg P_a(p)$	Remove b
12	$\neg P_{(a \vee b)}(p)$	Keep a and b
13	$\neg P_{(a \vee b)}(p)$ and $(a \subseteq b) \Rightarrow \neg P_a(p)$	Remove b
14	$\neg P_{(a \wedge b)}(p)$	Keep a and b
15	$\neg P_{(a \wedge b)}(p)$ and $(a \subseteq b) \Rightarrow \neg P_b(p)$	Remove a

Now consider §164.502(a)(5)(i) - Use and disclosure of genetic information for underwriting purposes:

> Notwithstanding any other provision of this subpart, a health plan, excluding an issuer of a long-term care policy falling within paragraph (1)(viii) of the definition of health plan, **shall not** use or disclose protected health information that is genetic information (GI) for underwriting purposes.

This section deals with a prohibition and includes four definitions. Here, only a subset of health plans are not allowed to use or disclose a subset of PHI, which is GI ($\neg P_a(u \vee d)$). Again, we have the case similar to rule 11. Thus, for simplification, we keep the subset of health plans that are not an issuer of a long-term care policy, and we keep GI and remove PHI.

Finally consider, §164.52(c) - Implementation specifications - Provision of notice:

> A covered entity **must** make the notice required by this section available on request to any person and to individuals as specified in paragraphs (c)(1) through (c)(3) of this section, as applicable.

In this case, we have an obligation statement and four definitions. Person subsumes individual ($Individual \subseteq Person$) and they have an "and" relationship, which is similar to rule 5: $O(Individual \wedge Person)$(make available to) and $(Individual \subseteq Person) \Rightarrow O_{person}$(make available to). Thus, we remove individual and reduce the number of definitions to three.

In this paper, we focused on removing the definitions with subsumption relationships via few examples. However, we need to validate the rules and their application both to the logic of these legal requirements and to their legal interpretation to determine the long-term viability of this proposed approach.

VI. THREAT TO VALIDITY

We now discuss the threats to validity of our work based on the roadmap presented by Perry et al. [19]. They define three types of validity as construct validity, internal validity and external validity.

Construct validity examines that to what extent the case studies actually measure answers to the questions. To mitigate this threat, we used grounded theory approach to build our ontology and started with incremental creation of the ontology from an empty set and classified definitions based on that set. We also moved back and forth between annotating the definitions in FBRAM and classification of definitions into categories and we revised the categories based on the new definitions and the annotated definitions.

Internal validity examines the bias and analyzes the causal relationships resulted from the data. To mitigate this threat, the authors of the papers checked the results separately and tried to reach a consensus on the characteristics of the definitions. Also, we built our method on the previously evaluated framework FBRAM.

External Validity verifies to what extent the results can be generalized. We applied our method to nine regulations in the healthcare domain in the US as part of a case study. While the results are still preliminary and only valid for this data set, we believe that we can reach saturation in our results and better generalize the result through additional case studies. We also extracted definitions based on a set of keywords such as defined as or synonymous to but as this list is not exhaustive, it is possible that we did not cover all of the definitions With a larger datasets in future, we can improve our keyword set to extract as many definitions as possible.

VII. CONCLUSION AND FUTURE WORK

In this paper, we used grounded theory approach to develop a method for analyzing and comparing the definitions from multiple jurisdictions. We analyzed 371 definitions from nine US regulations in the health care domain and developed an ontology for definitions. We extended FBRAM with new annotation codes to itemize the units of analysis in definitions and support easier comparisons. We defined heuristics and graphical models for comparison and showed how the graphical notation can help keep track of definitions in a legal statement. We also provided an analysis of how to integrate definitions in legal statement and how to simplify them.

In the future, we aim to extend the experiment to regulations in healthcare domain in other US states to reach saturation. Our ontology needs to be validated on a larger dataset to measure the generalizability of the categories. To date, we primarily focused on the *Information* category. However, as we saw in Section VI, definitions can refer to each other and we may find different results for other categories, such as technology. We also discussed the normalization of connectors in our models, as means to identify constraints and actors that should be compared in the graphical model. We aim to extend our work with the help of lexicons, such as Wordnet, and a tool support to automate the normalization process and help simplify the comparison process. We also propose a preliminary ap-

proach to simplify legal statements by unifying constraints in definitions to increase comprehensibility without losing legal coverage.

ACKNOWLEDGMENT

This work was partially funded by AFR-PDR grant #5810263 and the National Security Agency.

REFERENCES

[1] D. M. Blei, Probabilistic topic models, Commun. ACM, vol. 55, no. 4, pp. 77–84, Apr. 2012

[2] G. Boella, M. Martin, P. Rossi, L. van der Torre, and A. Violato, Eunomos, a legal document and knowledge management system for regulatory compliance, In ITAIS Conf., Springer 2012, pp. 571-578.

[3] T. D. Breaux, Exercising Due Diligence in Legal Requirements Acquisition: A Tool-supported, Frame-based Approach, In 17th IEEE International Requirements Engineering Conference (RE'09), Atlanta, Georgia, pp. 225-230, Sep. 2009.I.

[4] T. D. Breaux and A. I. Antón, A Systematic Method for Acquiring Regulatory Requirements: A Frame-Based Approach, In Proc. 6th International Workshop on Requirements for High Assurance Systems (RHAS-6), Delhi, India, Sep. 2007.

[5] T. D. Breaux, A. Antón, K. Boucher and M. Dorfman, Legal requirements, compliance and practice: An industry case study in accessibility, In 16th IEEE International Requirements Engineering Conference (RE'08), Barcelona, Spain, 2008, pp. 43-52.

[6] FrameNet, https://framenet.icsi.berkeley.edu/fndrupal/, Last accessed: June 2015.

[7] S. Ghanavati, D. Amyot, A. Rifaut and E. Dubois. Goal-Oriented Compliance with Multiple Regulations. In 22nd IEEE International Requirements Engineering Conference (RE'14), Karlskrona, Sweden, 2014

[8] S. Ghanavati, L. Humphreys, G. Boella, L. Di Caro, L. Robaldo, and L. van der Torre, Business Process Compliance with Multiple Regulations, 33th International Conference on Conceptual Modeling (ER'14), USA, 2014.

[9] D. G. Gordon and T. D. Breaux, Comparing requirements from multiple jurisdictions, In 4th RELAW, IEEE, 2011, pp. 43-49.

[10] D. G. Gordon and T. D. Breaux, Reconciling multi-jurisdictional legal requirements: A case study in requirements water marking, In 20th Int. Requirements Eng. Conf., IEEE CS, 2012, pp. 91-100.

[11] D. G. Gordon, and T. D. Breaux, Managing multi-jurisdictional requirements in the cloud: towards a computational legal landscape, In 3rd ACM Workshop on Cloud Computing Security, ACM, 2011, pp. 83-94.

[12] D. G. Gordon, T. D. Breaux. Assessing Regulatory Change through Legal Requirements Coverage Modeling, In 21st IEEE International Requirements Engineering Conference (RE'13), Rio de Janeiro, Brazil, pp. 145-154, Jul. 2013.

[13] B. G. Glaser, Conceptualization: On theory and theorizing using grounded theory, International Journal of Qualitative Methods, 2008

[14] S. Ingolfo, A. Siena and J. Mylopoulos, Goals and Compliance in Nomos 3, 7th International i* Workshop (iStar'14), Thessaloniki Greece. June 2014.

[15] S. Ingolfo, A. Siena and J. Mylopoulos, Modeling Regulatory Compliance in Requirements Engineering, 1st International Workshop on Conceptual Modeling in Requirements and Business Analysis (MReBA), Atlanta, U.S.A. October 2014.

[16] A. K. Massey, J. Eisenstein, A. I. Antón, and P. P. Swire, Automated Text Mining for Requirements Analysis of Policy Documents, In 21st IEEE International Requirements Engineering Conference, Rio de Janeiro, Brazil, July 2013.

[17] C. Matuszek, J. Cabral, M. Witbrock and J. DeOliveira, An introduction to the syntax and content of Cyc, AAAI Spring Symp. Formalizing and Compiling Bg. Knowledge and its Apps. to Knowledge. Rep. and Question Answering, pp. 44-49, 2006.

[18] J. C. Maxwell, A. I. Antón, P. P. Swire, M. Riaz and C. M. McCraw, A Legal Cross-References Taxonomy for Reasoning about Compliance Requirements, Requirements Engineering Journal, 2012.

[19] D. E. Perry, A. A. Porter, and L.G. Votta, Empirical studies of software engineering: a roadmap. In Proceedings of the Conference on the Future of Software Engineering (New York, NY, USA, 2000), ICSE '00, ACM, pp. 345–355.

[20] A. Siena, I. Jureta, S. Ingolfo, A. Susi, A. Perini and J. Mylopoulos, Capturing variability of law with Nómos 2, In Conceptual Modeling, vol. 7532 of LNCS, 2012, pp. 383-396.

[21] A. Siena, A. Perini, A. Susi and J. Mylopoulos, Towards a framework for law-compliant software requirements, In 31st Int. Conf. Software Engineering, IEEE CS, 2009, pp. 251-254.

[22] E. Rosch, Principles of categorization, Concepts: core readings (1999): 189-206.

[23] WordNet. http://wordnetweb.princeton.edu/perl/webwn, Last accessed: June 2015.