Mining Rule Semantics to Understand Legislative Compliance

Travis D. Breaux and Annie I. Antón
Department of Computer Science
North Carolina State University
{tdbreaux, aianton}@eos.ncsu.edu

ABSTRACT

Organizations in privacy-regulated industries (e.g. healthcare and financial institutions) face significant challenges when developing policies and systems that are properly aligned with relevant privacy legislation. We analyze privacy regulations derived from the Health Insurance Portability and Accountability Act (HIPAA) that affect information sharing practices and consumer privacy in healthcare systems. Our analysis shows specific natural language semantics that formally characterize rights, obligations, and the meaningful relationships between them required to build value into systems. Furthermore, we evaluate semantics for rules and constraints necessary to develop machine-enforceable policies that bridge between laws, policies, practices, and system requirements. We believe the results of our analysis will benefit legislators, regulators and policy and system developers by focusing their attention on natural language policy semantics that are implementable in software systems.

Categories and Subject Descriptors

D.2.1 [Requirements/ Specifications]: Languages and Methodologies

General Terms

Reliability, Security, Standardization, Languages, Legal Aspects

1. INTRODUCTION

Organizations seeking to achieve legislative compliance must (a) ensure that their company policies comply with legal regulations (e.g., standards and recommendations) and (b) guarantee that their business processes and operational systems implement their policies. Non-technical stakeholders (e.g. corporate or executive officers, policy analysts and lawyers) interpret these legal regulations in the context of their organizations and develop their organizational policies accordingly. In response, technical stakeholders (e.g. information technology (IT) managers and system administrators) interpret organizational policies to configure and deploy software systems that support the organization's overall business processes.

Deploying operational software systems that comply with legislation requires technical stakeholders to understand regulations in such a way that compliance can be guaranteed in the systems they implement. To this end, we are investigating and modeling regulatory semantics to support the development of a policy language that can codify law, policies, and system requirements that are properly aligned. Our prior work in analyzing Internet health care [1, 2] and financial [3] privacy policies provides a foundation for this work. These studies yielded over 1,200 unique, semi-structured goal statements that were extracted from over 100 Internet privacy policies using a technique called goal mining (the extraction of goal statements from texts during content analysis) [1, 4]. We then specified formal semantic models that distinguish goals as either rights or obligations [7, 8]; we refer to this specification process as *semantic parameterization*. *Rights* are activities that people or systems are permitted to do while *obligations* are activities that people and systems must do. Semantic models have properties that are desirable for comparing and disambiguating policy statements, regenerating natural language policy statements, and answering specific *what*, *where*, *how* and *why* questions.

Employing our experience in applying *Semantic Parameterization* to goal statements, in this paper we apply Semantic Parameterization to the Health Insurance Portability and Accountability Act (HIPAA) Fact Sheet [11], to develop formal rule semantics that can bridge the gap between natural language privacy policies and technical system policies. Furthermore, we validate our observations by cross-referencing our results with the HIPAA Privacy Rule [12] from which the Fact Sheet was originally derived. Our analysis reveals that certain keywords in regulatory text are indicative of compliance rules and constraints for both people and systems. In addition, we discuss the relevance of these semantics to existing privacy policy languages and note the importance of balancing specific rights with obligations to ensure that rights and obligations both have value. We believe the results of our analysis can help legislators, regulators and policy and software developers focus their attention on regulation and policy language with machine-enforceable semantics.

2. RELATED WORK

Several strategies have been proposed to derive formal models from the full scope of natural language (English), including conceptual dependencies [16] and conceptual graphs or semantic networks [18]. However, processing the full scope of natural language is excessive for analyzing privacy legislation. In privacy legislation, interesting natural language statements are limited to what tasks people and systems are entitled (rights and permissions) or obligated (responsibilities or requirements) to perform in order to satisfy legislative objectives. In addition to rules governing business processes, these statements include functional and non-functional system requirements. Within the limited scope of analyzing legislation, approaches have emphasized first-order logic models [5, 14, 15, 17].

First-order logic as a modeling notation provides sound and complete proofs of domain-specific properties. Generally, each variant of first-order logic provides certain benefits and limitations. In Section 5.2, we show that arithmetic operations are required to evaluate constraints from policy statements, yet these expressions are not decidable in first-order logic. However, it is still worth considering the strengths and weaknesses of logic-based models, since they uniquely describe the representational challenges to-date. Sergot et al. use deontic logic to model the

WPES'05, November 7, 2005, Alexandria, Virginia, USA.

British Nationality Act (BNA) of the United Kingdom [15]. Deontic logic provides semantics for describing rights and obligations [13]. They found that transcribing correct uses of negation from the BNA to logic were not straightforward and that counterfactual conditions within a single rule are prone to subjective interpretation. Sherman modeled the Canadian Income Tax Act in Prolog [17] in which he noted difficulty representing time and events in a model based on first-order logic. Sherman's model was also limited to absolute temporal relations between an event and a specific date and time. We show the additional need to specify time periods and relative temporal relations between events; both are specifications independent of specific calendar time. Alternatively, Antonious et al. explore the use of defeasible logic in analyzing and reasoning about regulations [5]. Defeasible logic allows prioritizing rules so that the highest priority rule fires unaffected by lower priority rules and they highlight its use in resolving logical inconsistencies. Finally, Kerrigan and Law describe the REGNET system developed for regulatory compliance assistance and tested in the domain of environmental law [14]. REGNET manages cross-references between regulation subtexts and XML associations between subtexts and simple logic rules. The system provides compliance assistance checking for consistency of logic rules across subtexts.

3. ANALYZING PRIVACY LEGISLATION

For this study, we analyzed the following HIPAA-related documents: 1) Fact Sheet: Protecting the Privacy of Patients' Health Information" [11] and 2) HIPAA Privacy Rule: Section 160, Subparts C, E and Section 162, Subparts E [12]. The Fact Sheet was prepared by the HHS to define rights and obligations established in the HIPAA Privacy Rule. The Fact Sheet is more amenable to analysis than the rule because it results from an effort that includes "answers to hundreds of common questions about the rule as well as explanations and descriptions about key elements of the rule" [11]. An important difference is the Fact Sheet excludes a complex matrix of cross-references distributed through the original Privacy Rule. Our analysis of the Privacy Rule shows a total of 439 cross-references across 22 sections of the rule with a maximum 71, mean 19.9, and median 17 cross-references. In addition, 38 cross-references referred to non-HIPAA documents. Each cross-reference qualifies the meaning of a statement by referring to a definition or statement in another section or less often in another document. Finally, the HIPAA Privacy Rule has eighteen times more words than the Fact Sheet, making the fact sheet a reasonable formative study.

The analysis procedure that we applied to the Fact Sheet is described in three steps which were repeated throughout that document: 1) identify a natural language statement that expresses rights, permissions, or obligations; next apply Semantic Parameterization to the statement to 2) derive semantic models for the actors, actions, and objects of each statement using the Knowledge Transformation Language (KTL) [7, 8] and 3) derive rules with pre-conditions and effects built from temporal constraints that relate semantic models. The two applications of Semantic Parameterization produce reusable natural language patterns making the process more consistent and repeatable.

Applying the Semantic Parameterization process to the entire Fact Sheet yielded encodings for 15 rights, 19 obligations and 12 rules. In addition, several reusable patterns were identified including seven patterns for rights, seven for obligations and nine for rules. These patterns are presented in Section 4. The process required 11 person hours; the first author spent only 4 hours initially with an additional 7 hours spent by both the first and second authors collaborating. Finally, we indexed the original Privacy Rule using the twenty-three patterns to validate our understanding of these patterns in specifying rights and obligations.

4. RIGHTS, OBLIGATIONS AND RULES

The natural language patterns correlate unique word sequences with specific parts of speech (e.g., modals, verbs, prepositions) to rights, obligations and constraints. The patterns for encoding rights and obligations are similar because they all identify a primary actor, action, and some relationship to other objects or activities. The patterns for encoding rules place constraints on rights and obligations and more frequently coincide with patterns for encoding obligations. In the following examples, an actor is either a provider of health-related services or products or a consumer or patient.

4.1. Patterns for Rights (R)

Rights define what an actor is allowed to do in terms of their capabilities. For example, an actor, such as a patient, may be capable of "seeing" their medical records however they may not have the expressed right to perform this activity. Table 1 shows the seven natural language patterns that were identified to consistently encode rights in the Fact Sheet and the number of times each pattern identified a right (***R***) in the Privacy Rule versus another semantic convention (***A***).

Table 1: Patterns for Rights

ID	*Pattern*	*R*	*A*
R_1	<actor> should/may be able to <verb> …	0	0
R_2	<actor> may <verb> …	119	17
R_3	<actor> can/could <verb> …	0	9
R_4	<policy> permits <actor> to <verb> …	3	1
R_5	<actor> would not have to <verb> …	0	0
R_6	<policy> does not restrict… <actor> …	0	0
R_7	<policy> does not require <actor> …	0	0

Among the seven patterns, three cases are highlighted. The most general case includes patterns R_1, R_2 and R_3 where a right to perform the action (a verb) is granted to a particular actor or group of actors using the modalities "should," "may," and "can." The patterns R_4, R_6 and R_7 demonstrate how the language used for obligations (see Section 4.2), such as "would have to," "restrict," or "require," are negated to establish a right. In other words, if an actor is not obligated to perform some action then they have the implied right to perform or not perform the action at their discretion. As we will see in Section 4.2, negating specific keywords for rights also establishes symmetric obligations.

4.2. Patterns for Obligations (O)

Obligations define the required behavior of an actor in one or more activities. Table 2 shows the seven patterns that were identified in the Fact Sheet that consistently encode obligations. The table also shows how often each pattern correctly identified an obligation (***O***) in the Privacy Rule as opposed to another semantic convention (***A***).

Table 2: Patterns for Obligations

ID	*Pattern*	*O*	*A*
O_1	<actor> should <verb> …	0	1
O_2	<actor> should be <verb'ed> …	0	0
O_3	<actor> will/would <verb> …	18	31
O_4	<actor> must/must be <verb'ed> …	189	0
O_5	<actor> which is charged with <verb'ing>	3	1
O_6	<policy> requires <actor> to <verb> …	1	0
O_7	<actor> may not <verb> …	30	0

The patterns for obligations have several notable characteristics. First, pattern O_1 and O_2 are similar except that the verb in O_2 is in the past-tense form and is preceded by the verb "be". It is foreseeable that pattern O_4 could have a similar relation with the modal "must" accompanied by a present-tense verb. Similar to the patterns for rights, the patterns for obligations include pattern O_6 that explicitly identifies the policy as the authority transferring obligations to actors. In addition, pattern O_7 uses the language for rights with negation to establish an obligation for the actor.

4.3. Patterns for Constraints (C)

For our purposes, rules associate pre-conditions with effects. Both pre-conditions and effects contain constraints that may describe activities, such as "a patient makes a request to a provider," or state such as "information is classified protected". If a pre-condition is true then a set of corresponding effects must also be true. From our analysis, pre-conditions and effects often included temporal constraints between the time of an activity and another activity or calendar time. Temporal constraints associate explicit times or sequences of events with activities and/ or states. In a rule, the pre-conditions will contain one or more conditions some of which describe activities, states and/ or have temporal constraints. The following nine constraint patterns C_1 through C_9 were extracted from the Fact Sheet.

Table 3: Patterns for Constraints

ID	*Pattern*
C_1	<actor> should be able to <verb>… if <actor/ object>… <verb>
C_2	<actor> may <verb>… but <actor> would not have to <verb>…
C_3	<actor> will <verb>… on/upon <event>…
C_4	<actor> may <verb>… for/for each <event>…
C_5	<actor> must <verb>… to ensure that <actor>… will <verb>…
C_6	<actor> would have to <verb>… before <verb>…
C_7	<actor> must first <verb>… before <verb>…
C_8	<actor> must <verb>… by <date>…
C_9	<actor> should <verb>… within <timeframe>…

There are two classes of rules distinguished by how the rule semantics exhibit temporal constraints. The first class has one event in the pre-conditions occurring before a second event in the effects as evidenced by patterns C_1 through C_7. In this case, the rule is equivalent to a temporal constraint between two activities. In patterns C_1 and C_2 the first activity is the effect for the second activity, the pre-condition. In patterns C_3 and C_4, the temporal constraints associated with the terms "on," "upon," "for," and "for each" establish the first activity as the effect of the latter activity, the pre-condition. In patterns C_5 through C_7, however, the first activity is the pre-condition for the latter activity, the effect.

5. ANALYSIS OF CONSTRAINTS

Applying Semantic Parameterization to the Fact Sheet yielded insights into natural language dynamics in policy statements that are required to specify rules with constraints in both pre-conditions and effects. We seek to generalize our observations and in particular we identify the need to represent cardinality, arithmetic operators, comparison relations, and ordinality as observed first in the Fact Sheet and later in the Privacy Rule.

5.1 Cardinality

Numbers in policy statements can be divided into two categories: symbolic and cardinal numbers. Symbolic numbers such as zip codes or social security numbers are strictly representational and may be treated as unique identifiers for a concept such as region or person, respectively. Cardinal numbers, however, specify a quantity of some concept. For example, the HIPAA Fact Sheet states several penalties (sanctions) for not complying with an obligation including fines from 100 dollars to 100,000 dollars and time in prison less than 10 years. In these cases, the concept is a penalty such as a fine in a number of dollars or a prison sentence in a number of years. In general, cardinal numbers always pair a numerical quantity, such as *100,000*, with a conceptual unit, such as *dollars*, and typically imply some named quantity such as a *fine*. Cardinal numbers are operands to arithmetic operators and comparison relations and may be used to establish an ordinal relation across a set of entities. In the Privacy Rule, we identified 64 different instances of cardinal numbers.

5.2 Arithmetic Operators, Comparison Relations

Natural language policy statements include adjectives (in inflected form) and prepositions that, used in conjunction with cardinal numbers or named quantities, indicate an increase or decrease (arithmetic) operation or a greater than or less than (comparison) relation. While a few keywords are generic such as more and less, most are relevant only to a specific named quantity. For example, the keywords *before*, *during*, and *after* are relevant to the named quantity *time*; *younger* and *older* are relevant to *age*; *shorter*, *longer*, *wider*, and *taller* are relevant to *width*, *height*, *length*, etc. In general, if the keyword is preceded by a numerical quantity and followed by a named entity with an implied reference to an appropriate named quantity, the keyword refers to an arithmetic operation. For example, a deadline described by the statement "30 days *after* the request" signals an arithmetic operation where "30 days" is added to "the time of request" to establish the deadline. Alternatively, the keywords may appear in a comparison relation between two entities. For example, in the Fact Sheet the statement "patients would have to sign a specific authorization *before* a covered entity could release their medical information" compares the time of two events establishing that one event occurs *before* another. Evidence for the use of these and similar keywords has been indexed in the HIPAA Privacy Rule and the results appear in Table 4 with totals for the number of instances that were arithmetic (***A***), comparative (***C***), and neither (***N***).

5.3 Ordinality

Ordinality refers to the index of an entity within an ordered set and is identifiable in HIPAA by adjectives including *first*, *second*, and *last*. Each index is relevant to contextual criteria. For example, the *first* event always refers to the *earliest* event in a set of events ordered by *time*. Ordinality depends on the existence of comparison relations to create an order, and as a result, the inflected-form adjectives each have an ordinal form that describes the *first* or *last* entity in an order. For example, the forms *least* and *most* are generic while others refer to specific concepts such as *earliest* and *latest* for *time*, *oldest* and *youngest* for *age*, *shortest*, *longest*, *widest* and *tallest* for *width*, *height*, *length*, etc. In the Privacy Rule, the ordinals *first* and *last* were most common with 7 and 3 occurrences, respectively. Typical usage for ordinals in the Rule include "the first disclosure during the accounting period" and the "individual's last known address."

Table 4: Inflected-form adjectives used in arithmetic, comparative operations.

Keyword	***A***	***C***	***N***	***HIPPA Privacy Rule Examples***
less	5	1	0	• not ***less*** than 30 days before… • ***less*** that 6 years from…
more	27	10	0	• no ***more*** frequently than once every… • contains ***more*** than 20,000 people…
before	1	9	9	• at least 15 days ***before*** the… • not less than 30 days ***before***…
after	20	8	2	• 180 days ***after*** the effective date… • ***after*** the compliance date…
older	0	1	0	• age 90 or ***older***…
smaller	0	1	0	• geographic subdivisions ***smaller*** than a state…
longer	2	7	0	• no ***longer*** than 30 days from the date… • no ***longer*** needed for the purpose…
during	12	4	0	• ***during*** the first year after… • ***during*** normal business hours…
within	25	0	5	• ***within*** 180 days of when… • ***within*** the time limit set…

6. DISCUSSION & SUMMARY

Our analysis revealed that in HIPAA constraints are often described using cardinal numbers, arithmetic operations, comparison relations and ordinals to distinguish entities in regulations. Understanding the relationship between constraints and the original regulation text is important in order to evaluate the ability of emerging policy languages to sufficiently express compliance requirements. For example, EPAL 1.1 [6] and Oasis XACML 2.0 [19] express numbers as attributes, however, they do not support the designation of units for these numbers — a source of potential ambiguity if one policy describes a time in minutes and another policy describes time in seconds. With regards to arithmetic operators and comparison relations, P3P [9] uses the W3C APPEL 1.0 Working Draft [10] for rules that lacks support for either, although support for comparison relations are declared as future work items. Alternatively, the EPAL 1.1 standard defers to XACML for conditions that allow both arithmetic operators and comparison relations. Finally, P3P, APPEL, EPAL and XACML all lack semantics to express ordinality over a set of related elements as observed directly in regulation texts.

Lastly, we observed that rights and obligations are complementary and that they must be balanced to ensure rights are both accountable and enforceable. For example, in the HIPAA Privacy Rule the patient may request that the healthcare provider restrict access to their protected health information, however, the provider is not obligated to honor that request [8]. Rights without complementary obligations are meaningless since governed parties are not required to respond to the invocation of such rights. In terms of designing and engineering software systems, these rights may be effectively ignored. On the other hand, obligations without an explicit and complementary right do have value and must be properly incorporated into system specifications.

7. REFERENCES

[1] A. I. Antón and J. B. Earp. "A Requirements Taxonomy to Reduce Website Privacy Vulnerabilities," Requirements Engineering Journal, Springer Verlag, 9(3), pp. 169-185, August 2004.

[2] A.I. Antón, J.B. Earp, M. Vail, N. Jain, C. Gheen and J. Frink. "An Analysis of Web Site Privacy Policy Evolution in the Presence of HIPAA," To appear in IEEE Security and Privacy, 2005.

[3] A.I. Antón, J.B. Earp, D. Bolchini, Q. He, C. Jensen and W. Stufflebeam, "The Lack of Clarity in Financial Privacy Policies and the Need for Standardization," IEEE Security & Privacy, v. 2 no. 2, pp. 36-45, 2004.

[4] A. I. Antón, J. B. Earp and A. Reese. "Analyzing Web Site Privacy Requirements Using a Privacy Goal Taxonomy," IEEE Joint Requirements Engineering Conference (RE'02), Essen, Germany, pp. 605-612, 9-13 September 2002.

[5] G. Antoniou, D. Billington and M. Maher. "On the Analysis of Regulations Using Defeasible Rules." *In Proc. of the AAAI-98 Workshop on Knowledge Management and Business Process Reengineering*, Madison, Wisconsin, pp. 46-50, July 1998.

[6] P. Ashley, S. Hada, G. Karjoth, C. Powers and M. Schunter. Enterprise Privacy Authoring Language (EPAL), version 1.1, http: //www.zurich.ibm.com/security/enterpriseprivacy/EPAL/Specification/

[7] T.D. Breaux and A.I. Antón. "Deriving Semantic Models from Privacy Policies." *In Proc. of the 6th Int'l Workshop on Distributed Systems and Networks (POLICY'05)*, Stockholm, Sweden, 6-8 June 2005, pp. 67 – 76.

[8] T.D. Breaux and A.I. Antón. "Analyzing Goal Semantics for Rights, Obligations, and Permissions." *In Proc. of the 13th Int'l Conf. on Requirements Eng. (RE'05)*, Paris, France, Aug. 29-Sept. 2, 2005.

[9] L. Cranor, M. Langheinrich, M. Marchiori, M. Presler-Marshall and J. Reagle. The Platform for Privacy Preferences (P3P), version 1.0, W3C Recommendation, http://www.w3.org/TR/P3P/

[10] L. Cranor, M. Langheinrich and M. Marchiori. A P3P Preference Exchange Language (APPEL), version 1.0. W3C Working Draft, http://www.w3.org/TR/P3P-preferences/

[11] "Fact Sheet: Protecting the Privacy of Patients' Health Information," published by the U.S. Department of Health and Human Services, Washington D.C., April 14, 2003.

[12] "Standards for Privacy of Individually Identifiable Health Information." 45 CFR Part 160, Part 164 Subpart E. In Federal Register, vol. 68, no. 34, February 20, 2003, pp. 8334 – 8381.

[13] A.J.I. Jones and M. Sergot. "Deontic Logic in the Representation of Law: Towards a Methodology." Artificial Intelligence and Law, Kluwer Academic Publishers, 1(1), pp. 45-64, March 1992.

[14] S. Kerrigan, K.H. Law. "Logic-based Regulation Compliance-assistance," *In Proc. of the Int'l. Conf. of the 9th Artificial Intelligence in Law (ICAIL'03)*, Edinburgh, Scotland, UK. pp. 126-135, June 2003.

[15] M.J. Sergot, F. Sadri, R.A. Kowalski, F. Kriwaczek, P. Hammond and H.T. Cory. "The British Nationality Act as a Logic Program" In Communications of the ACM, 29(5), pp. 370-386, May 1986.

[16] R.C. Shank. "Conceptual Dependency: A Theory of Natural Language Understanding," Cognitive Psychology, v. 3, no. 4, 1972, pp. 532 – 631.

[17] D. M. Sherman. "A Prolog model of the income tax act of Canada" *In Proc. of the 1st Int'l Conf. on Artificial Intelligence and Law (ICAIL-87)*, Boston, MA, USA, pp. 127-136, 1987.

[18] J.F. Sowa, Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, Reading, MA, 1984.

[19] T. Moses (ed.) eXtensible Access Control Markup Language (XACML), ver. 2.0 Oasis Standard. http://xml.coverpages.org/xacml.html