# Towards Natural-Language Understanding and Automated Enforcement of Privacy Rules and Regulations in the Cloud: Survey and Bibliography

Nick Papanikolaou, Siani Pearson, Marco Casassa Mont

Cloud and Security Lab
Hewlett-Packard Laboratories
{ nick.papanikolaou, siani.pearson, marco.casassa-mont }@hp.com

**Abstract.** In this paper we survey existing work on automatically processing legal, regulatory and other policy texts for the extraction and representation of privacy knowledge and rules. Our objective is to link and apply some of these techniques to policy enforcement and compliance, to provide a core means of achieving and maintaining customer privacy in an enterprise context, particularly where data is stored and processed in cloud data centres. We sketch our thoughts on how this might be done given the many different, but so far strictly distinct from one another, approaches to natural-language analysis of legal and other prescriptive texts, approaches to knowledge extraction, semantic representation, and automated enforcement of privacy rules.

**Keywords:** natural-language processing, privacy policies, policy enforcement, cloud computing

## 1 Introduction

Privacy is a central concern of growing importance in our society, especially with the continuing growth and popularity of online services that require the sharing of personal data. There have been numerous accounts of the notion of privacy and several attempts at formulating an adequate definition, but due to its complexity and dependence on context, this notion is highly elusive and difficult to pin down in such a way that satisfies the needs of customers and suppliers alike. In order to tackle citizens' privacy demands, numerous privacy laws and regulations have been enacted, and enterprises - which administer and manage huge databases of personally identifiable information as part of normal business practice - are required to demonstrate compliance with these laws and regulations. The consequences of failing to comply can be severe, not only for an enterprise but for its customers, whose ensuing loss of privacy could be very damaging to their reputations and finances (cf. identity theft). For this reason, there is a significant business need for tools and techniques to (automatically and efficiently) analyze privacy texts, extract rules and subsequently enforce them throughout the supply chain.

Clearly it is unreasonable to expect a computer program to fully understand a legal or other policy text. However there are a variety of techniques and programs for analyzing, annotating and extracting information from texts, and there have been various attempts at applying such techniques in the context of privacy. Furthermore there is much work on formalizing privacy and representing privacy-related properties in unambiguous logical form, which lends itself better to automated analysis. As for the enforcement of privacy rules and requirements, there exist rule-based systems that distil privacy knowledge into a form that can be executed directly by a machine. Of course, such systems have their limitations, but there are numerous practical benefits, especially as they reduce the effort required to ensure compliance significantly.

In the literature there is a tendency for the tasks and objectives in the previous paragraph (parsing and analysis of source texts, knowledge extraction, semantic representation, and enforcement of privacy rules) to be considered each in isolation. However these tasks may be seen as essentially interlinked processes in a privacy compliance lifecycle, so that the output of one task is the input of the next. This is a lifecycle because the processes need to be continually repeated in order to account for new privacy rules that emerge as societal needs, laws and regulations change. It should be noted here that there is some work being done in the European research project CONSEQUENCE to automate the full lifecycle involved in managing data sharing agreements, from capturing agreements in a pseudo-language to translating them into enforceable policies; however the mapping from legal texts to this pseudo-language representation is a human process and does not seem to involve automated natural-language analysis.

This paper attempts to survey work on natural language processing and semantic representation of legal and regulatory privacy-related knowledge. We have divided this survey into four principal groupings of papers:

1. Work on parsing and basic natural-language processing of legal and regulatory texts,
2. Work on knowledge extraction from texts,
3. Work on semantic representations of privacy and privacy-related knowledge, and
4. Work on automating compliance with privacy legislation and regulation.

It is our interest to see how these techniques can be combined and cross-linked in order to build a platform for automatically recognizing and enforcing privacy rules.

## 2 Parsing and Basic Analysis of Source Texts

First we consider practical approaches – namely, tools and techniques - to the task of natural-language processing of legal and regulatory texts. It is worthwhile to note that the main techniques for analysis of such texts tend to have many similarities across different domains, whether the texts refer to healthcare regulations, business best practice, or privacy rules: the common element is the prescriptive nature of the texts. In particular, texts that consist exclusively of detailed descriptions of rules often use standard sentence structures and patterns, which can be identified and formalized to a significant degree. What changes with different application domains, naturally, is the vocabulary, and the frequency of particular word clusters (see [11] for statistical results regarding the vocabulary and phrases common to privacy policies in particular). The papers [1], [2], [14], [15] all describe different tools for analysis of prescriptive texts, and we review these next. Also of note is work in the IBM REALM project [24].

Moulin and Rousseau [1] describe a "knowledge acquisition system" known as SACD, implemented in Prolog, which has been developed for the analysis of regulatory texts. SACD was used to process the National Building Code of Canada, so that its stipulations could be represented in a machine-readable, indexable and searchable form. What is particularly interesting about SACD is that it can adapt the knowledge representation structures it uses automatically, as it processes input. Furthermore it provides a graphical user interface during the process of syntactic analysis, which allows for user intervention when a particular text fragment has been decomposed incorrectly or inaccurately. SACD makes use of chart-parsing algorithms and Prolog definite-clause grammars, which are ideally suited to the low-level analysis of sentence structure and meaning. The authors do mention the fundamental limitation of their approach, namely that the built-in grammars need to be repeatedly revised and extended to be able to parse new language elements, vocabulary and usage patterns. However, the system is capable of adapting its representations of knowledge, as text is parsed. Moulin and Rousseau's paper describes work that falls into almost all of the categories in our classification, including knowledge representation and learning; the link to compliance is mentioned, but the authors do not explain how it might be automatically achieved using their tool.

Michael, Ong and Rowe [2] and Ong [15] develop an architecture and concrete tool for analyzing texts with prescriptive rules. Their approach alludes to knowledge extraction and logical representation of rules, but the tool they present is specifically an extractor which turns a prescriptive sentence into a 'meaning list'; how this meaning list can be used by a handler or enforcement mechanism is beyond the scope of their work. In terms of textual analysis, their approach is to use an off-the-shelf part of speech (POS) tagger and to process its output to determine whether the input sentence describes an obligation, permission or interdiction; the meaning list resulting from analysis of the sentence identifies the different actors involved and their interrelationships.

Brodie et al. [14] present a tool, SPARCLE, designed for authoring technically enforceable privacy policies using natural language. The tool is designed to parse and interpret English text describing privacy rules, and generate from that text appropriate XACML policies. In particular, the tool provides a structured policy authoring environment. What is appealing about SPARCLE is the ability to link access control statements to the original natural language requirements; this aids both understanding and transparency.

## 3 Knowledge Extraction from Texts and Learning

Antón and a number of different collaborators (see [3,5,6,7,8]) have used textual mining techniques to analyze privacy policies and a number of different privacy and privacy-related regulations. For example, in [3] the authors focus on privacy policies from financial institutions which claim to be compliant with the Gramm-Leach-Bliley Act (GLBA). Papers [3] and [5] refer to the use of a tool called PGMT (Privacy Goal

Management Tool), which is a tool for representing and analyzing rules arising in privacy regulations as restricted natural-language statements. In [6] the authors discuss the extraction of structured rules from source texts using an NLP platform called *Cerno*.

In [7] Breaux, Anton and Vail use their approach of *semantic parameterization* to represent the US HIPAA (Health Insurance Portability and Accountability Act) Privacy Rule as a set of restricted natural-language statements, classified as rights, constraints or obligations. They identify standard phrases appearing in the legislative document, and note the frequency of their occurrence and the corresponding modality (right/obligation/interdiction etc.). They also discuss how to handle ambiguities. This work is extended further in [8], where the authors develop a detailed classification of constraints and introduce means of handling complex cross-references arising in the legal text of the HIPAA.

Delannoy et al. [9] combine a template-matching technique with machine learning in order to match rules from the Canadian 1991 tax guide with text describing case studies of particular individuals; this approach in principle allows one to see which tax rules apply in a given situation. The paper describes an architecture and tool called MaLTe, which is capable of learning how to apply rules to different input texts.

Delisle et al. [10] describe in detail a framework for extracting meaning from the structure of technical documents. Their approach is relevant to the analysis of prescriptive texts in that they assume that input documents are highly structured and somewhat predictable. The authors propose a number of techniques for identifying patterns in texts and converting sentences to Horn clauses. The Horn clauses represent knowledge about the domain in question; through the use of machine learning techniques, this knowledge is extended and refined as more documents are supplied.

Stamey and Rossi [11] use singular-value decomposition and latent semantic analysis techniques to analyze privacy policy texts. They identify commonly occurring topics and key terms and their relations. They are also able to detect similar word meanings; the strength of their approach is that they are able to pick out ambiguities in privacy policies and make them visible to the user. The tool *Hermes* developed by the authors allows automated analysis of an entire privacy policy text, outputting an overall ranking of the policy (when compared to a reference text).

We are also aware of much work on knowledge extraction from legislation [16-20]. Due to space limitations we will not expand on this further.

## 4 Semantic Models and Representations

Waterman [4] develops a simple table-based representation of particular laws. This author demonstrates a so-called ‘intermediate isomorphic representation’ of a rule from the US Privacy Act, and similarly for a rule from the Massachusetts Criminal Offender Records Law. The key idea here is to use a structured representation that can be mapped directly back to the original legal text and to corresponding computer code. The representation still uses natural language, but with additional logical structure. The additional structure helps to separate out actors, verbs, context and particular constraints that exist in the legal text (and which are often implied or included indirectly with the use of cross-references).

The framework proposed by Barth, Datta, Mitchell and Nissenbaum [12] comprises a formal model which is used to express and reason about norms of transmission of personal information. This work does not involve automatically analyzing text, but does provide a formalism for manually representing notions of privacy found in legislation – particularly in the texts of HIPAA, COPPA and GLBA. The formal model provides notations for defining sets of agents communicating via messages, with particular roles, in specified contexts; linear temporal logic, with a past operator, allows one to express properties that the agent behaviours should satisfy. Policy compliance is formally defined in terms of this model. Although the authors assume that their formalism is for a human user, we envisage the possibility that using natural-language analysis it should be possible to extract from texts some privacy rules expressed in this formalism.

May, Gunter and Lee [13] define a semantic model for expressing privacy properties, and apply it to the HIPAA Privacy Rule; it is based on a classical access control model used in operating system design. The authors translate the legal text into a structured format that uses the commands in the proposed access control model to express rules. The paper does not restrict itself to representation; once the legal rules have been formally expressed, the authors use the SPIN model checker [23] to automatically reason about the consistency of the generated rules; they demonstrate subtle differences between the year 2000 and year 2003 versions of the HIPAA Privacy Rule.

It is clear that a uniform, consistent, formal representation of privacy knowledge and privacy rules in particular is useful for automated reasoning about privacy issues. We are keen to make use of existing formal representations of privacy rules when performing natural-language analysis of privacy-related texts, since the usefulness of such representations has already been demonstrated for complex texts, particularly American privacy legislation.

## 5 Policy Enforcement and Compliance

We are not aware of any previous work that addresses the whole lifecycle of natural-language analysis of privacy texts with the goal of enforcing suitable rules, e.g. in an enterprise setting (although the EU CONSEQUENCE project mentioned before does take an holistic approach it does not involve natural-language analysis). As stated in the Introduction, achieving compliance with privacy legislation and regulations is a central concern in enterprises, and means of automating compliance are highly desirable. Since new privacy rules are almost exclusively expressed using natural-language, means of automatically analyzing the appropriate texts and extracting rules from them necessary – the resulting rules can then be incorporated into existing enterprise rule-bases, such as those used in compliance checkers or information governance (GRC) platforms. We mention here some work on automated policy enforcement and compliance, which has so far been developed separately and independently of any consideration of automated knowledge and rule extraction.

The EnCoRe research project [24] is developing a platform for expressing and enforcing privacy preferences for personal data; recent case studies include a system for managing data held within an enterprise's HR systems, and health data stored about individuals and tissue samples in a biobank. Through the use of a suitable policy enforcement architecture, legal and regulatory privacy rules, along with individuals' privacy preferences, can be automatically enforced so that unauthorized and/or unsuitable access to data is prevented. In [21] we proposed a simple conceptual model for representing privacy rules, which can be directly mapped to technically enforceable access control policies (expressed e.g. using XACML).

In [22] Pearson et al. propose a tool for providing decision support with regards to privacy-sensitive projects that arise in an enterprise. Decision support systems are built on knowledge bases with rich sets of rules, and the process of translating legal texts, regulations and corporate guidelines into technically enforceable rules is complex and laborious. For this reason a conceptual model is a useful aid.

There is much work on aspect-oriented access control for privacy [25,26]. Also, Peleg et al. [27] have proposed a framework for situation-based access control that is useful for handling the privacy of patients' health records. Bussard and Becker have extended their previous work on formalising access control policies to privacy-related scenarios in [28].

We believe there is scope for integration of several of the different approaches described so far into a natural-language processing pipeline, which can be integrated with technical enforcement mechanisms to achieve compliance for privacy: this starts with the initial task of analyzing natural-language privacy texts, to the extraction of formalized rules and their automatic enforcement.

We are working on developing tools for automating privacy in cloud computing and, for this, natural-language analysis of provider SLAs, international laws and regulations will need to be combined with suitable enforcement methods such as distributed access control [29], sticky policies and policy-based obfuscation [30].

## 6 Review and Future Work

We have in this paper surveyed a number of existing works related to the analysis of privacy and privacy-related texts, with the goal of representing the knowledge and rules therein in a logical form that is machine readable and automatically enforceable. We presented a grouping of these works in four classes, which may be seen as constituting essential steps in a natural-language processing pipeline; such a pipeline may be seen as a workflow that would be included as part of the compliance lifecycle for an enterprise.

We have focused specifically on research in the privacy space, due to its growing significance as a societal concern and the critical consequences faced by enterprises that fail to meet compliance with privacy law and regulations.

## References

1. B. Moulin and D. Rousseau. Automated Knowledge Acquisition from Regulatory Texts. *IEEE Expert* **7**(5), 27--35 (2002)
2. J. Bret Michael, V. Ong, and N. C. Rowe. Natural-Language Processing Support for Developing Policy-Governed Software Systems. In *Proceedings of 39th International Conference and Exhibition on Technology of Object-Oriented Languages and Systems (TOOLS 39)*, 263--274 (2001)
3. A. Antón, J. B. Earp, Q. He, W. Stufflebeam, D. Bolchini, C. Jensen. Financial Privacy Policies and the Need for Standardization. *IEEE Security and Privacy* **2**(2), pp. 36--45 (2004)
4. K. Krasnow Waterman. Pre-processing Legal Text: Policy Parsing and Isomorphic Intermediate Representation. In *Proceedings of PRIVACY 2010 - Intelligent Information Privacy Management AAAI Spring Symposium*, Stanford Center for Computers and Law, Palo Alto, California, USA (2010)
5. T. D. Breaux and A. I. Antón. Deriving Semantic Models from Privacy Policies. In *Proceedings of the Sixth International Workshop on Policies for Distributed Systems and Networks (POLICY'05)* (2005)
6. N. Kiyavitskaya, N. Zeni, T. D. Breaux, A. I. Antón, J. R. Cordy, L. Mich, J. Mylopoulos. Extracting Rights and Obligations from Regulations: Toward a Tool-Supported Process. In *Proceedings of ASE'07* (2007)
7. T. D. Breaux, M. W. Vail and A. I. Antón. Towards Regulatory Compliance: Extracting Rights and Obligations to Align Requirements with Regulations. In *Proceedings of 14th IEEE International Requirements Engineering Conference (RE'06)* (2006).
8. T. D. Breaux, A. I. Antón. Analyzing Regulatory Rules for Privacy and Security Requirements. *IEEE Transactions on Software Engineering* **34**(1), 5--20 (2008)
9. J. F. Delannoy, C. Feng, S. Matwin, and S. Szpakowicz. Knowledge Extraction from Text: Machine Learning for Text-to-rule Translation. In *Proceedings of Machine Learning and Text Analysis Workshop (ECML-93)* (1993)
10. S. Delisle, K. Barker, J. Delannoy, S. Matwin, S. Szpakowicz. *From Text to Horn Clauses: Combining Linguistic Analysis and Machine Learning*. In *Proceedings of Canadian AI Conference (AI/GI/CV '94* (1994)
11. J. W. Stamey, R. A. Rossi. Automatically Identifying Relations in Privacy Policies. In *Proceedings of SIGDOC'09* (2009)
12. A. Barth, A. Datta, J. C. Mitchell, H. Nissenbaum. Privacy and Contextual Integrity: Framework and Applications. In *Proceedings of IEEE Symposium on Security and Privacy* (2006)
13. M. J. May, C. A. Gunter, I. Lee. Privacy APIs: Access Control Techniques to Analyze and Verify Legal Privacy Policies. In *Proceedings of Computer Security Foundations Workshop (CSFW'06)* (2006)
14. C. A. Brodie, C. Karat, J. Karat. An Empirical Study of Natural Language Parsing of Privacy Policy Rules Using the SPARCLE Policy Workbench. In *Proceedings of Symposium on Usable Privacy and Security (SOUPS)* (2006)

15.V. L. Ong. *An Architecture and Prototype System for Automatically Processing Natural-Language Statements of Policy*. Master's thesis, Naval Postgraduate School, Monterey, California (2001)

16.J. Davies, M. Grobelnik, D. Mladenic (eds.). *Semantic Knowledge Management*. Springer (2009)

17.J. Breuker, P. Casanovas, M. C. A. Klein, E. Francesconi (eds.). *Law, Ontologies and the Semantic Web*. IOS Press (2009)

18.P. Casanovas, G. Sartor, N. Casellas, R. Rubino (eds.). *Computable Models of the Law*. Springer (2008)

19.V. R. Benjamins, P. Casanovas, J. Breuker, A. Gangemi (eds.). *Law and the Semantic Web*. Springer (2005)

20.Danièle Bourcier. *Legal Knowledge and Information Systems*. IOS Press (2003)

21.Marco Casassa Mont, Siani Pearson, Sadie Creese, Michael Goldsmith, Nick Papanikolaou. A Conceptual Model for Privacy Policies with Consent and Revocation Requirements. In *Proceedings of PrimeLife/IFIP Summer School 2010: Privacy and Identity Management for Life*, Lecture Notes in Computer Science, Springer (2010)

22.S. Pearson, P. Rao, T. Sander, A. Parry, A. Paull, S. Patruni, V. Dandamudi-Ratnakar, P. Sharma. Scalable, accountable privacy management for large organizations. In *Proceedings of 13th Enterprise Distributed Object Computing Conference Workshop (EDOCW 2009)*, 168--175 (2009)

23.SPIN. http://www.spinroot.org

24. IBM REALM Project. http://www.zurich.ibm.com/security/publications/2006/REALM-atIRIS2006-20060217.pdf

25.K. Chen, D. Wang. An aspect-oriented approach to privacy-aware access control. In *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007* (2007)

26.C. Vanden Berghe, M. Schunter. Privacy Injector - Automated Privacy Enforcement through Aspects. In *Proceedings of 6th Workshop on Privacy Enhancing Technologies, 28-30 June 2006* (2006)

27.M. Peleg, D. Beimel, D. Dori, Y. Denekamp. Situation-Based Access Control: privacy management via modeling of patient data access scenarios. *Journal of Biomedical Informatics* (to appear).

28.L. Bussard, M. Y. Becker. Can Access Control be Extended to Deal with Data Handling in Privacy Scenarios? In *Proceedings of W3C Workshop on Access Control Application Scenarios* (2009).

29.M. Y. Becker, P. Sewell. Cassandra: Distributed Access Control Policies with Tunable Expressiveness. In *Proceedings of 5th IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY 2004), 7-9 June 2004, Yorktown Heights, NY, USA*. 159--168, IEEE Computer Society (2004)

30. M. Mowbray, S. Pearson and Y. Shen. Enhancing privacy in cloud computing via policy-based obfuscation. *Journal of Supercomputing*. DOI: 10.1007/s11227-010-0425-z.