A Requirements-based Comparison of Privacy Taxonomies

Aaron K. Massey and Annie I. Antón
Department of Computer Science, North Carolina State University
{akmassey, aianton}@ncsu.edu

Abstract

Understanding the nature of privacy regulation is a challenge that requirements engineers face when building software systems in financial, healthcare, government, or other sensitive industries. Requirements engineers have begun to model privacy requirements based on taxonomic classifications of privacy. Independently, legal research has modeled privacy harms in a taxonomic fashion. In this paper, we compare a requirements engineering taxonomy of privacy protections and vulnerabilities to a legal taxonomy of privacy harms. We seek to determine the extent to which the concepts and terminology are consistent between the two taxonomies. A consistent, standard vocabulary for privacy concepts for both requirements engineers and lawyers will improve the common understanding of privacy concepts, legal traceability and compliance auditing. We conclude that the taxonomies we analyzed are reasonably compatible. We believe this compatibility indicates that a taxonomic understanding of privacy is a promising area of research for requirements engineers.

1. Introduction

Regulatory compliance has been the primary driver of information security in industry since 2005 [9]. Ernst and Young's 2006 annual survey of almost 1,300 organizations found that "privacy and data protection" was the key motivating factor for information security practices in 42% of the organizations [8]. In 2007, this figure increased to 58% [9]. However, past research has noted several reasons why it is difficult to build software systems that comply with regulations [7, 13, 11]. In particular, complying with privacy and data protection policies is especially challenging [3, 6, 12].

The ability to build a system that complies with privacy legislation depends greatly on the ability to understand the meaning of privacy. Engineers responsible for building such systems must parse the subtle and difficult-to-define concepts found in privacy legislation and policy. These engineers also need to be able to reason effectively about vulnerability scenarios relating to the same legislation and policy to ensure proper coverage of the software requirements. The requirements generated for these purposes must be clear enough to enable software engineers to implement functional software that complies with regulations and legislation in a cost effective and timely manner. Each of these steps in the construction of a new information system is a challenge.

This paper compares and contrasts two separate taxonomies created to improve the understanding of privacy. If these taxonomies can ensure a common vocabulary between engineers and lawyers, then they could play a key role in establishing legal compliance. The first is the Antón-Earp requirements taxonomy, which was introduced at RE'02. This taxonomy is designed to increase a software engineer's understanding of privacy related software requirements based on the pre-requirement goals extracted from an organization's online privacy policies [4, 2]. The second is a legal taxonomy of privacy harms, which we will refer to as the Solove Taxonomy. The Solove Taxonomy is designed to provide a comprehensive understanding of the plurality of privacy problems recognized by cultures around the world [14, 15]. Because of the nature of their development, the Antón-Earp Taxonomy has a narrower scope.

This paper analyzes the extent to which the vocabulary used to describe privacy harms and vulnerabilities overlaps between the legal and requirements engineering communities, respectively. We hypothesized that the Antón-Earp Taxonomy of vulnerabilities would be a subset of the Solove Taxonomy. Ideally, if the vocabulary were conceptually consistent, it would be possible to map each vulnerability unambiguously from the Antón-Earp Taxonomy to a distinct privacy harm found in the Solove Taxonomy.

The remainder of this paper is structured as follows: Section 2 introduces both taxonomies. Section 3 compares the taxonomies by examining their similarities and differences. Section 4 discusses conclusions that can be drawn from a requirements engineering perspective. Finally, Section 5 highlights needed future work.

2. Background

This section introduces the both privacy taxonomies.

2.1. The Antón-Earp Taxonomy

The Antón-Earp Taxonomy was developed using grounded theory to analyze 25 e-commerce privacy policies during the summer of 2000 [2, 10]. The grounded theory was manifest in the form of goal

2008 Requirements Engineering and Law (RELAW'08)
978-0-7695-3630-9/08 $25.00 © 2008 IEEE

mining — a repeatable technique supplemented by a set of heuristics for deriving structured natural language pre-requirement goals from text artifacts [1, 5, 4]. Applying grounded theory to online privacy policies revealed 12 categories of privacy elements spread across two broad classifications as shown in table 1 below [4, 2].

Table 1: The Antón-Earp Taxonomy

Protection Goals	*Vulnerabilities*
Notice/Awareness	Information Monitoring
Choice/Consent	Information Aggregation
Access/Participation	Information Storage
Integrity/Security	Information Transfer
Enforcement/Redress	Information Collection
	Information Personalization
	Solicitation

Privacy protection goals safeguard the privacy of a customer's data and there are five categories as follows: Notice and Awareness goals describe how a customer is informed about an organization's practices regarding their data. Choice and Consent goals describe a customer's ability to choose how they want their data to be managed by an organization. Access and Participation reflects a customer's ability to challenge, correct or modify their data as used by an organization. Integrity and Security goals describe measures an organization takes to protect the accuracy and security of a customer's data. Enforcement and Redress goals describe the ways that organization approaches internal policy violations by their employees.

Vulnerabilities reflect a potential privacy violation and there are seven categories as follows: Information Monitoring describes how an organization tracks customers' interaction with their website. Information Aggregation reflects the ways that an organization will combine customer data with third-party data sources. Information Storage reflects an organization's practices regarding what/how customer records are stored in the organization's database. Information Transfer describes how an organization may share their collected customer information with affiliates and third-parties. Information Collection shows what types of information an organization may collect and how that organization collects the specified information. Information Personalization reflects the methods an organization uses to tailor the presentation of their website to their customers. Solicitation shows the purposes and methods an organization would use to contact their customers.

2.2. The Solove Taxonomy

The Solove Taxonomy is based on an attempt to conceptualize the social and legal aspects of privacy from the bottom-up rather than define privacy as a singular concept from the top-down [15]. Originally described in a law review article, the Solove Taxonomy has been revised and highlighted as a way to understand privacy [14, 15]. The goal of this taxonomy is improve privacy legislation and policy by dividing the concept of privacy into discrete, actionable elements [15]. The taxonomy is divided into 16 categories spread across four broad classifications as shown in table 2 below [15].

Table 2: The Solove Taxonomy

Information Collection	*Information Processing*
Surveillance	Aggregation
Interrogation	Identification
	Insecurity
Information Dissemination	Secondary Use
Breach of Confidentiality	Exclusion
Disclosure	
Exposure	*Invasion*
Increased Availability	Intrusion
Blackmail	Decisional Interference
Appropriation	
Distortion	

Information Collection deals exclusively with privacy problems resulting from gathering information. Surveillance consists of methods of watching, listening and recording a subject's activities. Interrogation describes methods an organization may use to ask or elicit information from a subject.

Information Processing describes methods to store, modify or manipulate a subject's information. Aggregation combines individual and previously separate pieces of data about a subject. Identification depicts an organization's methods for determining which individual is described by a set of data. Insecurity is a failure to properly protect stored data. Secondary Use reflects the use of data for a purpose other than that for which it was originally provided. Exclusion describes the inability of a subject to have knowledge of how their data is being used.

Information Dissemination consists of privacy harms resulting from the release of information about a subject. Breach of Confidentiality contains those harms based on the violation of a trust agreement to maintain confidentiality of a subject's information. Disclosure describes harms related to the release of truthful

978-0-7695-3630-9/08 $25.00 © 2008 IEEE

information about a data subject. Exposure describes the dissemination of information about a subject's grief, body or bodily functions. Increased Accessibility consists of the ways that a subject's public information may be made available to a wider audience than before. Blackmail involves a threat made to a data subject about a potential release of their information. Appropriation describes the use of a subject's identity or information to serve the purposes of the organization rather than the subject. Distortion consists of harms related to the release of falsified information about a data subject.

Invasion consists of the various intrusions on an individual's private life. Intrusion is a form of invasion that describes all harms resulting from the disturbance of an individual's peace and solitude. Decisional Interference is an invasion into a subject's decisions about their private affairs.

3. Taxonomies Compared

This section compares and contrasts the Antón-Earp Taxonomy with the Solove Taxonomy to better understand the extent to which it is possible to map the classifications across both taxonomies.

The Antón-Earp Taxonomy is split between one classification that describes measures to prevent harms and another classification that describes measures that could lead to privacy harms. The Solove Taxonomy classifies only privacy harms because Solove's goal is to outline all possible privacy harms. Obviously, it would not be possible to list all possible mechanisms by which privacy may be protected. Because the protection goals describe a protective action, we will focus exclusively on mapping the vulnerabilities outlined in the Antón-Earp Taxonomy to privacy harms in the Solove Taxonomy.

Our methodology of comparison consisted of comparing each vulnerability from the Antón-Earp Taxonomy to each privacy harm category in the Solove Taxonomy and determining if the vulnerability could reasonably be interpreted as being a subset, a superset or completely unrelated. Of the seven vulnerabilities in the Antón-Earp Taxonomy, three vulnerabilities each map as proper subsets of three different categories from the Solove Taxonomy.

The Information Monitoring vulnerability in the Antón-Earp Taxonomy is a proper subset of the Surveillance privacy harm. Antón and Earp describe Information Monitoring vulnerabilities as "information tracking by organizations when consumers visit their Web site" [2]. Solove describes the Surveillance privacy harm as "the watching, listening to, or recording of an individual's activities" [15]. This definition fits clearly as a proper subset of the Surveillance privacy harm.

Information Aggregation vulnerabilities are a proper subset of the Aggregation privacy harm in the Solove Taxonomy — they are described in nearly identical terms. Antón and Earp describe Information Aggregation as "refer[ing] to combining previously gathered [personally identifiable information] data with data from other sources" [2]. The phrase "previously gathered" synchronizes nicely with Solove's characterization of Aggregation as falling under the Information Processing classification. Solove describes the Aggregation privacy harm as "involv[ing] the combination of various pieces of data about a person" [15]. Clearly, these classifications are quite similar.

Information Transfer vulnerabilities are a proper subset of the Disclosure privacy harm; these two categories are also described in strikingly similar terms. Antón and Earp describe Information Transfer as "the practice of allowing information to be transmitted, the reason(s) why information may be transferred, and to whom that information is transferred" [2]. Solove defines Disclosure as "occur[ring] when certain true information about a person is revealed to others" [15]. Both descriptions are extremely similar and focus specifically on the transfer of information rather than its eventual use.

The remaining four vulnerabilities from the Antón-Earp Taxonomy do not clearly map as proper subsets of a single privacy harm category in the Solove Taxonomy. These four vulnerabilities are areas of ambiguity that must be clarified for requirements engineers and lawyers to build a common understanding of potential privacy problems.

Despite the fact that four of the seven vulnerabilities could not be mapped to single privacy harms, the degree of ambiguity in the mapping of Antón-Earp vulnerabilities to Solove harms may not be insurmountable for a requirements engineer. Three of the four ambiguous vulnerabilities can clearly be mapped to a combination of two specific harms found in Solove's Taxonomy. The remaining vulnerability could map to a combination of three harms. Thus, even in the worst case an Antón-Earp vulnerability can be seen as unrelated to 13 of the 16 harms found in the Solove Taxonomy. Depending on the circumstances, this may clarify the situation well enough to make a legal compliance or engineering decision.

We now compare each of the four ambiguous Antón-Earp vulnerabilities with the set of privacy harms from the Solove Taxonomy to which they may correspond. Information Storage vulnerabilities may map to Surveillance or to Insecurity privacy harms depending on the nature of the goal. Antón and Earp describe information Storage as both "how and what records are stored in an organization's database" [2]. Obviously, the storing of records allows an organization the ability to perform surveillance as defined by Solove previously. However, Solove also defines Insecurity as a privacy harm where

978-0-7695-3630-9/08 $25.00 © 2008 IEEE

"carelessness in protecting stored information from leaks and improper access" [15]. If the method of storage used by an organization is insufficient, then an Insecurity privacy harm would be the result.

Information Collection vulnerabilities may map to Surveillance, Secondary Use or Breach of Confidentiality privacy harms. Antón and Earp describe Information Collection as "reflect[ing] what information is collected by Web sites" [2]. Some Information Collection vulnerabilities are potential Surveillance harms. However, some Information Collection vulnerabilities allow a third party to collect information without the customer's knowledge. These vulnerabilities would thus map to Secondary Use, which Solove defines as "the use of information for a purpose different from the purpose for which it was collected without the data subject's consent" [15]. Lastly, third party collection within Information Collection vulnerabilities could lead to Breach of Confidentiality privacy harms, which Solove defines as "breaking a promise to keep a person's information confidential" [15].

Information Personalization vulnerabilities may map to either Identification or Appropriation privacy harms. Antón and Earp describe Information Personalization as "the tailoring or customization of a Web site to a specific visitor, thus affecting the functionality or content offered to individual visitors" [2]. For example, recognizing a returning customer to change the web site's appearance involves identifying the customer's records in the organizations' database. Solove's Identification privacy harm is described as "linking information to particular individuals," which would have to occur for the site to be customized [15]. However, if a web site is tailored to someone by including advertising based on purchases made by the target's friend found in a social network, then that could be classified as appropriating the friend's identity. Appropriation is described by Solove as "the use of a data subject's identity to serve another's aims and interests" [15].

Solicitation vulnerabilities can map to Interrogation or Secondary Use privacy harms. Antón and Earp define Solicitation as "how and for what purpose organizations contact visitors or others" [2]. When an organization contacts an individual to obtain feedback regarding their own products, this is an Interrogation privacy harm, which Solove defines as "the pressuring of individuals to divulge information" [15]. However, some Solicitation vulnerabilities expressly allow third parties to contact customers to advertise products — these are clearly Secondary Use privacy harms as previously defined by Solove.

4. Conclusions

The mapping of vulnerabilities to privacy harm categories was, on the whole, not flagrantly ambiguous. In fact, six of the seven vulnerabilities in the Antón-Earp Taxonomy mapped to at most two categories of privacy harms. This similarity suggests that the two taxonomies are reasonably compatible.

Previously we hypothesized that the Antón-Earp Taxonomy would be a subset of the Solove Taxonomy, which we have shown to be true. However, the Antón-Earp Taxonomy has seven vulnerabilities that cover privacy harms from 10 of the 16 categories in the Solove Taxonomy. Such coverage is an interesting result and may be attributed to the nature of the two taxonomies. The Antón-Earp Taxonomy was built using a goal-based content analysis of 25 privacy policies for online information systems based in the United States. The Solove Taxonomy was built as an attempt to analyze all possible privacy harms in all cultural environments and may concern itself with items that are irrelevant to an information system or to the privacy culture in the United States. Both taxonomies describe privacy concerns that may be present, but do not necessarily have to occur in a given set of circumstances. However, the scope os Solove's study is clearly broader in nature.

The key conceptual difference between the two taxonomies is that of describing a goal as opposed to a harm. Because Solove's Taxonomy does not describe goals to achieve or maintain, any engineer attempting to build a system based on that terminology would have to define their own fitness and acceptance criteria. In contrast, the Antón-Earp Taxonomy emphasizes concerns that must be considered to increase requirements coverage and reduce vulnerabilities in web-based information systems.

This key difference is a serious concern for legal compliance in systems which must comply with privacy legislation. Without specific goals, the privacy requirements in legislation are challenging to meet from a compliance standpoint. In this paper we have shown that the two taxonomies are reasonably compatible. This finding suggests that the Antón-Earp Taxonomy could be useful in translating legislation that can be classified by the Solove Taxonomy into specific goals. Thus, by sequentially "chaining" these taxonomies, it may be possible to generate software requirements methodologically from privacy legislation.

More broadly, the independent use of privacy taxonomies in both the legal and requirements engineering communities suggests that both fields have found value in considering privacy as a plurality of discrete elements rather than a single uniform concept. The structure of a taxonomy allows for improved management and evolution of a topic. Additionally, a taxonomy can help ensure consistent application of responses to particular classifications of problems. The use of taxonomies in both law and engineering suggests that these benefits apply to both fields.

2008 Requirements Engineering and Law (RELAW'08)
978-0-7695-3630-9/08 $25.00 © 2008 IEEE

5. Future Work

We now discuss future work needed to improve the understanding of privacy requirements and communication between requirements engineers and lawyers. For the sake of simplicity and space limitations, we did not consider priorities and exceptions in our comparison of these taxonomies. On a given topic, real-world legislation is typically filled with case law and cross-references that further clarify the exceptions and priorities in written law. These clarifications, priorities and exceptions may affect the taxonomic understanding of privacy. As a result, determining how to model priorities and exceptions for legal requirements is an area for future research.

We have also refrained from analyzing the Solove Taxonomy as a stakeholder document to be used in the construction of requirements for an information system. Furthermore, Solove's Taxonomy was created with the goal of affecting the understanding of privacy in future privacy legislation. Requirements engineers may wish to study both of these to gain a better understanding of how useful Solove's Taxonomy is for requirements engineers building systems that must comply with privacy legislation.

Another promising area of future research includes attempting to build a privacy legislation compliant system based on chaining the two taxonomies as described in section 4. For example, the Solove Taxonomy could be used to classify the legislation into specific harms. Then, those harms could be mapped to Antón-Earp vulnerabilities. Finally, those vulnerabilities can be used to generate maintenance and avoidance goals and requirements. Each element of this process could be recorded to maintain traceability and leverage the benefits of easier maintenance provided by the use of a taxonomy.

The nature of the Antón-Earp Taxonomy is targeted narrowly to the specific privacy policies analyzed. It could be expanded through analysis of additional privacy policies. These additional policies could continue to be web site policies, but there is no reason that the methodology used could not be expanded to desktop or standalone applications as well. Expansion could supplement the Antón-Earp Taxonomy through the addition of vulnerabilities from the Solove Taxonomy's list of potential privacy harms which would result in a broader range of legislation for which compliance could be made easier. In addition, the vulnerabilities in the Antón-Earp Taxonomy found to map to multiple harm categories in the Solove Taxonomy could be disambiguated so that a clear one-to-one mapping is created.

6. Acknowledgements

This work was supported by NSF Information Technology Research Grant #522931.

7. References

[1] A. Antón. Goal-based requirements analysis. *Proceedings of the Second International Conference on Requirements Engineering*, Colorado Spring, CO, pages 136–144, 15-18 April 1996.

[2] A. I. Antón and J. B. Earp. A requirements taxonomy for reducing web site privacy vulnerabilities. *Requirements Engineering*, 9(3):169–185, 2004.

[3] A. I. Antón, J. B. Earp, and R. A. Carter. Precluding incongruous behavior by aligning software requirements with security and privacy policies. *Information and Software Technology*, 45(14): 967–977, 2003.

[4] A. Anton, J. Earp, and A. Reese. Analyzing website privacy requirements using a privacy goal taxonomy. *Proceedings of the IEEE Joint International Conference on Requirements Engineering*, pages 23–31, 2002.

[5] A. Antón and C. Potts. The use of goals to surface requirements for evolving systems. *Proceedings of the 1998 International Conference on Software Engineering*, pages 157–166, 19-25 April 1998.

[6] T.D. Breaux and A.I. Antón. Analyzing Regulatory Rules for Privacy and Security Requirements. *IEEE Transactions on Software Engineering,* 34(1), pp. 5-20, January 2008.

[7] T. D. Breaux, M. W. Vail, and A. I. Antón. Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations. *14th IEEE International Requirements Engineering Conference (RE'06)*, pages 49–58, Washington, DC, USA, September 2006. IEEE Society Press.

[8] Ernst & Young. "2006 Global Information Security Survey," November 2006.

[9] Ernst & Young. "10th Annual Global Information Security Survey," December 2007.

[10] B. Glaser and A. Strauss. *The Discovery of Grounded Theory.* Aldine, Chicago. 1967.

[11] G. T. Lau, S. Kerrigan, K. H. Law, and G. Wiederhold. An e-government information architecture for regulation analysis and compliance assistance. *In ICEC '04: Proceedings of the 6th international conference on Electronic commerce*, pages 461–470, New York, NY, USA, 2004. ACM.

[12] F.Massacci, M.Prest, N.Zannone. "Using a Security Requirements Engineering Methodology in Practice: The compliance with the Italian Data Protection Legislation," Technical Report DIT-04-103, 2004.

[13] P.N. Otto and A.I. Antón. Addressing Legal Requirements in Requirements Engineering. *15th IEEE International Requirements Engineering Conference*, pp. 5 - 14, 15-19 October 2007.

[14] D. Solove. "A Taxonomy of Privacy," *University of Pennsylvania Law Review*, Vol. 154, No. 3, pp. 477 - 560, January 2006.

[15] D. J. Solove. *Understanding Privacy*. Harvard University Press, 2008.