A Data Purpose Case Study of Privacy Policies

Jaspreet Bhatia and Travis D. Breaux
Institute for Software Research, Carnegie Mellon University
Pittsburgh, Pennsylvania, United States
{jbhatia, breaux}@cs.cmu.edu

Abstract — **Privacy laws and international privacy standards require that companies collect only the data they have a stated purpose for, called collection limitation. Furthermore, these regimes prescribe that companies will not use data for purposes other than the purposes for which they were collected, called use limitation, except for legal purposes and when the user provides consent. To help companies write better privacy requirements that embody the use limitations and collection limitation principles, we conducted a case study to identify how purpose is expressed among five privacy policies from the shopping domain. Using content analysis, we discovered six exclusive data purpose categories. In addition, we observed natural language patterns to express purpose. Finally, we found that data purpose specificity varies with the specificity of information type descriptions. We believe this taxonomy and the patterns can help policy analysts discover missing or underspecified purposes to better comply with the collection and use limitation principles.**

Index Terms — **data purpose, information types, natural language processing, privacy, policy, content analysis, requirements.**

I. Introduction

A large number of people in America use the Internet [16], and are thus impacted by the data practices of the website companies. Government and standards organizations thus provide regulatory guidance to help protect individual privacy. The Organization of Economic Cooperation and Development (OECD) privacy framework introduces the *data quality principle*, which states that "personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes." In addition, the OECD *purpose specification principle* requires that data purposes, which are the purposes for which data will be used, be specified no later than at the time of data collection. According to the *use limitation principle*, personal information should not be shared or used for previously unspecified purposes, except for when required by law, or by the consent of the subject [14].

In the U.S. and Europe, the OECD guidelines on purpose specification and use limitation appear in standards and other guidance. The U.S. National Strategy for Trusted Identities in Cyberspace[1] describes the *data minimization principle*, which states that organizations should only collect personally identifiable information to achieve specified purposes and only retain such information for the time needed to fulfill those purposes. The U.S. Circular A-130 and the NIST Privacy Engineering Framework both recognize the OECD guidelines, including purpose specification and use limitation, as a means to protect personal information. In Europe, the General Data Protection Directive (GDPR), Article 5, requires that data only be processed for explicit and specified legitimate purposes, and never in a way that is incompatible for those purposes. Analyzing data purposes does not only help comply with standards and laws, but also helps users better understand the data practices of the website companies, and helps them make informed decisions about using services on the Internet.

We conducted a case study to understand how data purpose is expressed by companies in their privacy policies. Most large companies in the U.S. and Europe have privacy policies, which are also required by certain laws. While privacy policies are difficult for the average person to read [13], they are also required by the U.S. Federal Trade Commission and the GDPR to be accurate and truthful. In the case study, we sought to: 1) identify and categorize data purposes, 2) study the variations in data purposes due to the variations in information types, and 3) analyze the scenarios in which data purpose could be inferred from information actions and information types.

The remainder of this paper is organized as follows: in Section II, we review the related work; in Section III, we present our approach to identify and categorize data purposes in privacy policies using content analysis; in Section IV, we present results; and in Section V, we discuss the future work.

II. Related Work

We now review prior work related to data purpose, including the Platform for Privacy Preferences (P3P) standard and the role of purpose in privacy requirements,

Cranor et al. conducted privacy surveys to determine the aspects of privacy policies that would likely be of most interest to users. From these surveys, they found that data purpose for which user data would be used was one among the three most important areas, the other two being, type of data collected, and whether or not data would be shared [9]. This motivated our first research question which aims to categorize data purposes in privacy policies into different categories. This would help users better understand the types of purposes for which their information could be used.

The P3P 1.0 Specification [10], which supports the *purpose specification principle* [14], defines a P3P vocabulary that includes eight major components, one of which is the "purpose" component, which concerns how collected data is

[1] https://obamawhitehouse.archives.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf

2332-6441/17 $31.00 © 2017 IEEE
DOI 10.1109/RE.2017.56

used, and whether individuals can opt-in or opt-out of any of these uses. The specification defines eleven purpose sub-elements, each representing a data use. In addition, each of these purpose sub-elements has a "required" attribute that indicates whether the data may be used for this purpose all the time, on an opt-in basis, or on an opt-out basis [10].

He and Antón propose a framework for modeling privacy requirements, where *purpose binding* is an important privacy requirement [12]. Purpose binding is that the data collected for one purpose should not be used for another purpose without user consent, and it can be modeled as permission constraints in the framework. They model the relationships between purposes using a purpose hierarchy. If an operation is allowed for a given purpose, it is also allowed for all sub-purposes.

Antón and Earp suggest three privacy evaluation criteria which could be used for rating policies, one of which is, whether a site contacts visitors for purposes beyond the primary purpose of data collection [1]. They also describe the *choice and consent goal* to ensure that the consumers are given the option to decide what personal information collected about them is to be used and whether it may be used for secondary purposes. In the *integrity and security goals*, the authors suggest that organizational procedures be used to limit access and avoid unauthorized purposes. Using collected user data for secondary purposes is a potential privacy violation where the consumer is acutely aware of or which they eventually become aware. The *contact goals* in their *privacy vulnerability taxonomy* deal with how and for what purpose organizations contact consumers using their personally identifiable information. This inspired our third research question about how data purposes vary with information types to add to the analysis suggested by Anton and Earp.

Privacy policies often place restrictions on the purposes for which an entity may use user data. Tschantz et al. propose a semantics of purpose restrictions which could be used as a formal or automated method for enforcing privacy policies [18]. Their formalism determines whether an action is for a purpose or not and is based on planning, using a modified version of Markov Decision Processes (MDPs). In their model, an action is for a purpose if and only if the action is part of a plan for optimizing the satisfaction of that purpose under the MDP model. They provide an auditing algorithm based on their formalism. In requirements engineering, Breaux et al. introduced the Eddy language for expressing and detecting conflicts among privacy requirements, which consist of a data action and data purpose, using Description Logic [6, 5]. A later extension to the Eddy language allowed checking policies for compliance with the collection and use limitation principles based on purpose specifications [8]. Bhatia et al. developed a semi-automated framework to extract privacy goals from privacy policies [3]. This approach can be extended to extract the data purposes which can be checked for compliance using the Eddy language. To automatically extract data purposes from privacy policies we would need to first understand how these data purposes are expressed syntactically in the policies, which motivated our second research question, "How are data purposes expressed in privacy policies?"

III. Analyzing Data Purposes

We now introduce our research questions and case study design based on content analysis. Our research questions are:

RQ1. What are the different data purpose categories associated with data practices?

RQ2. How are data purposes expressed in privacy policies?

RQ3. How do data purposes vary with the associated information types, within and across policies?

The case study design consists of policy sampling criteria and policy analysis method. For the policy sample, we chose a convenience sample of five policies from the shopping domain (see Table I). These policies are from shopping companies who maintain both online and "brick-and-mortar" stores.

TABLE I. Privacy Policy Dataset For Purpose Study

Company Name	Last Updated
Amazon	06/12/2012
Barnes and Noble	05/07/2013
Costco	12/31/2013
Lowes	04/25/2015
Walmart	9/17/2013

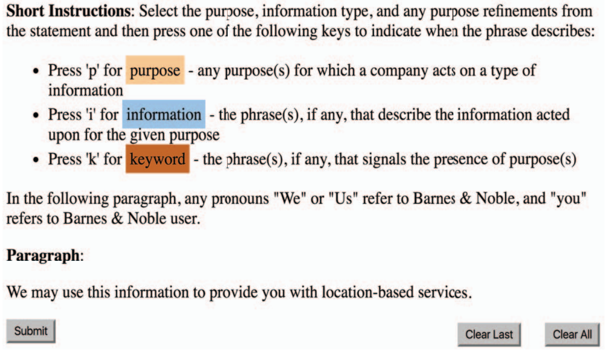
We first annotated the five policies to identify the data purposes, information types and the keywords, before applying content analysis [17] as follows: (1) we applied open-coding to categorize the annotated data purposes; (2) we reviewed the purposes and organized them by natural language lexico-syntactic features to discover patterns for expressing purpose; and (3) we classified the annotated data purposes as either *ambiguous* or *unambiguous*, to decide whether data purpose specificity varies with information type. Manual annotation is commonly used to establish a gold standard for evaluating unsupervised learning, such as clustering, which we plan to explore in future work.

The first step to answering our research questions is to use first-cycle coding [17] to extract the data purposes, the information actions which are associated with the data purposes, and the keyword that signals the presence of the purpose. We use the extracted data purposes and information actions, for all our studies, which we designed to answer our research questions. We now describe this step.

A. Extracting Data Purposes and Information Types

We manually annotated the privacy policies to identify the data purposes, corresponding information types and the keyword(s). We limited this analysis to statements about collection, use, disclosure and retention, which map to statements that Antón and Earp used to extract information collection, information transfer and information storage goals as part of the Privacy Vulnerability Taxonomy [1].

Consistent with the pre-processing described by Breaux and Schaub [7], the five policies are first prepared by removing section headers and boilerplate language that does not describe relevant data practices (e.g., table of contents). Next, the policy text is divided into ~120 word paragraphs to reduce fatigue during annotation. Finally, the paragraphs are collected in an input file for an Amazon Mechanical Turk (AMT) task. The task employs an annotation tool [7], which allows annotators to select relevant phrases matching a category, in this case, the data purpose, the information type and the keyword as shown in Figure 1. The first author performed this annotation task.



Short Instructions: Select the purpose, information type, and any purpose refinements from the statement and then press one of the following keys to indicate when the phrase describes:

- Press 'p' for purpose - any purpose(s) for which a company acts on a type of information
- Press 'i' for information - the phrase(s), if any, that describe the information acted upon for the given purpose
- Press 'k' for keyword - the phrase(s), if any, that signals the presence of purpose(s)

In the following paragraph, any pronouns "We" or "Us" refer to Barnes & Noble, and "you" refers to Barnes & Noble user.

Paragraph:

We may use this information to provide you with location-based services.

Fig. 1. Task to annotate data purposes

We now discuss the steps to answer our research questions.

B. Study Designs to Answer Research Questions

Research question RQ1 asks, “what are the different data purpose categories associated with data practices in privacy policies?” To answer this question, the annotator uses second cycle coding [17] to categorize the annotated data purposes into distinct categories. During the categorization task, the annotator also develops guidelines to determine the inclusion and exclusion criteria. These guidelines help to define the category, and ensure that categories remain exclusive and non-overlapping.

Research question RQ2, asks “how are data purposes expressed in privacy policies?” We used open coding [17] to identify and categorize the natural language patterns which are used to express purposes in the fives policies in our dataset. We categorized the patterns used to express data purposes based on their syntactic structure. The syntactic structure is a relative ordering of the data purpose (DP), the corresponding information type (IT) and the keyword (K) which signals the presence of the data purpose. For example, consider the statement from the Costco privacy policy: “We may provide to a third-party information as necessary to fulfill an order you have placed with us…” In this statement, the information type “information” precedes the keyword “as necessary to,” which is followed by the purpose “fulfill an order you have placed with us.” Thus, the syntactic structure used to express this data purpose is IT-K-DP. Similarly, consider the statement from the Amazon privacy policy “to help us make e-mails more useful and interesting, we often receive a confirmation when you open e-mail from Amazon.com.” Herein, the purpose “make e-mails more useful and interesting” is preceded by the keyword “to help us,” which is followed by the information type “confirmation when you open e-mail from Amazon.com.” The syntactic structure used to express this data purpose is K-DP-IT.

Research question RQ3, asks “how do data purposes vary with information types, within and across policies?” One way that data purposes and information types vary is by the extent to which the descriptions of each are imprecise [4]. To answer this question, the first author labeled each annotated information type as either ambiguous or unambiguous. For example, the information types “personal information” and “contact information” were categorized as ambiguous, whereas the information types “name” and “IP address” were annotated as unambiguous. While ambiguity is a matter of degree, the purpose of this categorization is to strictly delineate between extreme values. This categorization was also motivated by Evans et al., wherein the authors constructed an ontology from hyponyms, which are more specific terms, and hypernyms, which are more general terms, that were linked by keywords (“such as,” or “for example”) [11]. In general, abstract information types are annotated as ambiguous, whereas more specific types are annotated as unambiguous.

We adopt a similar bifurcation for data purposes. We annotate purposes with a broad interpretation as ambiguous, and narrowly described purposes as unambiguous. For example, an ambiguous purpose in the Barnes and Noble Policy reads, “provide you with a superior customer experience and, as necessary, to administer our business,” and an unambiguous purpose reads, “create features like Top Sellers,” wherein Top Sellers is a website feature name.

While annotating the information types and data purposes, the annotator develops a few guidelines. For example, for anaphora in information types (e.g., “this information), the annotator should resolve the anaphora to determine the information type associated with the data purpose, and then annotate the resolved information type instead. For example, in the statement from the Barnes and Noble policy, “…may use this information to fulfill your order and for other purposes…”, the information type “this information” refers to “name, email address, IP address, and shipping or billing address,” from a previous statement. Therefore, we annotate this referenced list and classify the list as unambiguous.

After classifying data purposes and types as ambiguous or unambiguous, we next compared the frequencies of each classification using the chi-squared test to determine if there was a correlation between these variations [15]. The chi-squared test uses the contingency table frequencies to determine the relationship, if any, between two variables.

IV. Results

The analysis of the five privacy policies (see Table I) produced 218 data purpose annotations. The annotations are classified into six categories as follows:

- *Service Purpose (SP)* – any purpose for which a company acts on a type of information to provide services to the user, including advertisements, preference-based content and improving the website services. For e.g., to provide services to the user, and to personalize the user services.
- *Legal Purpose (LP)* – any legal purpose for which a company acts on a type of information, including court orders, regulatory purposes or for any other legal reasons. For e.g. to comply with court order and to comply with regulatory requirements.
- *Communication Purpose (CP)* – any purpose for which a company acts on a type of information to communicate with the user about products, services and user issues, and to

provide user with notifications and updates. For e.g., to contact user to resolve issues, respond to user queries and update about new products.

- *Protection Purpose (PP)* – any user-data protection and fraud detection purpose for which a company acts on a type of information. For e.g., to detect data being matched to a machine and to make sure data was converted using an authorized machine.
- *Merger Purpose (MP)* – any purpose for which a company acts on a type of information in case of mergers, transfer of control, or transfer of company assets. For e.g., for merger negotiations and sales of company assets.
- *Vague Purpose (VP)* – any vague purpose for which a company acts on a type of information, the reason or consequence of which is unclear. For e.g., for any other use deemed helpful and for emergency purposes.

Table II shows the frequency of the different data purposes per category for each of the five policies.

TABLE II. FREQUENCY OF DATA PURPOSES ACROSS POLICIES

Privacy Policy	SP	LP	CP	PP	MP	VP	Total
Amazon	40	3	6	5	0	1	55
Barnes & Noble	34	3	4	0	2	4	47
Costco	21	4	5	0	1	0	31
Lowes	26	2	6	2	0	2	38
Walmart	29	6	3	8	1	0	47
Total	**150**	**18**	**24**	**15**	**4**	**7**	**218**

Table III presents examples for each data purpose category, along with the frequency that purposes appeared in the category across all the five policies in our dataset.

The most frequent category was service purpose (150/218 purposes) which includes a broad range of purposes: providing services to the user, advertising and marketing, improving the functionality of the website, personalizing the content to users' preferences, and analyzing users' data, among others. In contrast, legal purposes are very similar, for example, law enforcement purposes, and enforcing terms and conditions.

The research question RQ2 asks, "how are data purposes expressed in privacy policies?" Recall from Section III.B that we categorized the patterns used to express data purposes based on their syntactic structure. The syntactic structure is a relative ordering of the data purpose (DP), the corresponding information type (IT) and the keyword (K) which signals the presence of the data purpose. See Section III.B for examples.

The resulting syntactic categories are as follows:

- *DP* – The statement only contains the data purpose.
- *K-DP* – A keyword precedes a data purpose, and an information type is either missing or present in a previous statement. When the information type is missing, a policy often states that the company uses a technology to perform the data purpose, however it remains unclear which information type is associated with the data purpose.
- *IT-K-DP* – An information type precedes a keyword, followed by a data purpose.
- *K-DP-IT* – A keyword precedes the purpose, followed by an information type. This category is rare.
- *K-IT-K-DP* – A keyword precedes an information type, which is followed by another keyword, followed by the data purpose. In this category, the first keyword that signals the presence of a data purpose is the action "use."

TABLE III. EXAMPLES OF DATA PURPOSES

Data Purpose Category	Example Purposes	% Freq.
Service Purpose (SP)	location-based services, such as advertising, search results, and other personalized content (Amazon); we may customize our home page for you, better display pages according to your browser type (Costco)	68.8%
Legal Purpose (LP)	comply with the law (Amazon); enforce or apply our Conditions of Use and other agreements (Amazon)	8.3%
Communication Purpose (CP)	communicate with you about special offers, promotions, and other marketing programs and news that may be of interest to you (Barnes and Noble); process, evaluate and respond to your requests, inquiries and applications (Lowes)	11%
Protection Purpose (PP)	security and operational purposes, such as to measure traffic patterns (Walmart); help prevent and detect fraud and to offer certain credit or financial services (Amazon)	6.9%
Merger Purpose (MP)	in connection with a merger or sale involving all or part of Walmart or as part of a corporate reorganization or stock sale or other change in corporate control (Walmart); if Barnes & Noble becomes involved in a merger, acquisition, restructuring, reorganization, or any form of sale or other disposition of some or all of its assets (Barnes and Noble)	1.8%
Vague Purpose (VP)	other purposes (Lowes); own purposes (Lowes)	3.2%

In Table IV we present the keywords we identified for each syntactic category, and the corresponding frequency of the category across the five policies in our dataset. For the category K-IT-K-DP, we show the two keywords separated by a hyphen. In addition to the syntactic categories, 22/218 data purposes were implied from information actions, which we further discuss in Section V.A.

The research question RQ3 asks, "how do data purposes vary with information types?" Table V presents a contingency table for data purposes (DP) and information types (IT) that were classified as ambiguous (A), or unambiguous (U).

We calculated the Pearson's chi-squared statistic ($\chi^2(1)$=13.74, p=0.0002) from the contingency table in Table V. For degrees of freedom=1 and p=0.0002, the critical value is less than the chi-squared statistic $\chi^2(1)$>10.83. [2]. Thus, we conclude that the variations in the two variables IT and DP have a statistically significant correlation. Notably, an ambiguous information type (A-IT) is more likely to be associated with an unambiguous data purpose (U-DP). For example, in the Amazon's privacy policy, the information type "information" was associated with specific purposes, such as "location based services, advertising, etc." For unambiguous information types (U-IT), there is an almost equally likely chance that the IT will be linked to an unambiguous or ambiguous data purpose. For example, the Barnes and Noble privacy policy states that they use the types "name, email address, IP address, and shipping or billing address" for U-DP "to fulfill your order," in addition to the A-DP "for other purposes." To maximize customer awareness, one may expect to see more cases where U-IT is associated with U-DP. For example, the Costco privacy policy is where the information type "purchase information from our Health Care Centers" is associated with the unambiguous purpose "calculate your Executive Member 2% Reward." In contrast, an undesirable case would be where A-DP is associated with a A-IT, for example, in the Amazon privacy policy, A-DP "perform their functions" is associated with A-IT "personal information."

TABLE IV. KEYWORD TAXONOMY FOR SYNTACTIC CATEGORIES

Category	Refinement Keywords	Freq.
DP	for example, examples include	43
K-DP	allow us to, to help us, is necessary to, our purpose is to, so that we can	10
IT-K-DP	needed to, in an effort to, used to, use to, to, is appropriate to, for, helps us, help us, needed to, so that they may, is necessary to, allows us to, in order to, for purposes including, to provide, that provide, so that we can, that, to allow, for use in providing, offer you, provided to, in response to, for any purpose other than, will be providing, enable us to, to help	101
K-DP-IT	to help us	2
K-IT-K-DP	use - to provide you with, use - for such purposes, use - for, use - in order to, use - to	40

DP: Data Purpose, IT: Information Type, K: Keyword

TABLE V. CONTINGENCY TABLE FOR INFORMATION TYPES AND PURPOSE VARIATIONS

Information Type (IT)	Data Purpose (DP)		
	A-DP	U-DP	Total
A-IT	17	49	66
U-IT	105	93	198
Total	122	142	264

A: Ambiguous, U: Unambiguous, DP: Data Purpose, IT: Information Type

V. FUTURE WORK

We now discuss our future work, in light of the observations that we made during our case study.

A. Implied Purposes from Information Actions

We observed that some data purposes could be inferred from information actions. For example, consider the following statement from the Amazon policy: "You provide most such information when you search, buy, bid, post, participate in a contest or questionnaire, or communicate with customer service." In this statement, the user provides their information when they are using the website's "search" functionality. Thus, one can infer that the collected information is used for search purposes. The merger purpose category contains purposes that can be inferred from merger-related actions, as shown in the Barnes and Noble policy statement "… your data may be transferred to or shared with a third party as part of a sale, merger, or acquisition of Barnes & Noble…" In future work, we will investigate the relationship between data action and implied purpose, and whether data actions are typically linked to broader purpose categories, e.g., searches, likes or page views may be linked to marketing purposes, since these actions are the ones more likely surveilled to discover user interests.

B. Implied Purposes from Information Types

We also found that purpose could be inferred from information types, based on the functions that the types are used to perform, or the applications the types are used in. For example, the information type, "email address," can be used for the function "contacting the user," whereas the user can be contacted for different applications, such as for "marketing" or "transactions." Among the 305 unique information types extracted from our dataset of five policies, 205 were associated with either implied or explicitly stated data purposes, whereas the other 100 information types were not associated with any data purpose. Example information types not linked to a data purpose include: "1-click settings," "custom content," and "firmware version." We observed information types among these 100 types that were indirectly linked to purpose through hypernymy relationships, such as those hypernyms discovered by Evans et al. [11]. For example, the information type "usage information" was not found to implicitly associate with a data purpose. Surprisingly, "usage information" was defined as a subtype of "personal information" in the Barnes and Noble privacy policy, which in turn was associated with data purposes "to provide superior customer experience."

C. Third Party Purposes

We also observed that data purpose can be either a first- or third-party purpose. For example, Costco states that "We use this information for system administration…" Here, the data purpose is fulfilled by the company with whom the user does business, therefore it is a *first-party purpose*. On the other hand, Walmart states that it discloses personal information to third parties, so that third parties can "help with business operations" and "provide services on Walmart's behalf." Herein, the user's information is acted upon by third parties to fulfill business purposes, which is a *third-party purpose*.

Among the 218 purposes, 183 purposes are first-party data purposes, and the remaining 35 data purposes (Amazon-10, Barnes and Noble-10, Costco-7, Lowes-2, Walmart-6) are third-party purposes. Among the 35 third-party purposes, there was one purpose in the Walmart policy in which the purpose could be performed by either the first- or the third-party.

In future work, we plan to examine third party purposes in more detail to examine which kinds of data are shared with third parties, and under what specific purposes, and to measure the degree to which those purposes are ambiguous.

D. Hypernymy in Data Purposes

We also found multiple instances of data purpose hypernymy, wherein abstract data purposes were refined into specific data purposes. For example, the Lowes privacy policy defines "data analytics and system administration purposes, such as to determine whether you've visited us before or are new to the Site." In this statement, the purposes "data purpose and system administration purposes" are the abstract purposes, called hypernym purposes, which are refined into a more specific purpose “to determine whether you've visited us before” and “[to determine whether you] are new to the Site,” which are called hyponyms. We identified only five hypernymy relationships among purposes. Such instances of data purpose hypernymy, can be in used in future to construct a hierarchy of purposes. This hierarchy can be used to infer when permitting or prohibiting a generic purpose entails (through the hierarchy) a corresponding permission or prohibition of more specific purposes that are subordinate to the general purpose.

E. Examining Purpose Across Domains

While our case study examined the shopping domain, which covers a wide range of online activity by users, we envision creating a larger dataset of data purposes by covering a larger number of domains, such as telecommunications, health, employment, and social networks. This larger dataset would permit unsupervised learning and help companies and users understand how data purpose varies across different policies within and across domains. We believe that our analysis results could help us better correlate information types with data purposes, thus identifying the default or most probable purposes associated with each information type. Given a new policy, we could then determine whether the policy describes an expected purpose for a given information type, if an unexpected purpose was present, or if an expected purpose was missing.

F. Data Purposes Spanning Multiple Statements

In our analysis, we observed multiple instances where the data purposes, information type, and the keyword span multiple statements. For example, the Walmart policy states: “we may disclose your information in other special circumstances. These include situations when the sharing is necessary to protect the safety, property, or other rights of Walmart, our customers, our associates, or any other person, or where otherwise required by law. Examples include protecting the health or safety of customers…” In this paragraph, the information type “information” from the first statement is being used for the purposes mentioned in the second and third statements. The keyword “necessary to” signals the presence of protection purpose in the second statement. The third statement provides specific example purposes for the more generic purpose of protecting the customers. In future work, we envision the need to develop techniques that can trace purposes and their corresponding information types across multiple statements in order to reconstruct the context of data use.

Acknowledgment

We thank the CMU RE Lab for their helpful feedback. Funded provided by NSF Frontier Award #1330596, NSF CAREER Award #1453139 and NSA #141333.

References

[1] A.I. Antón, J.B. Earp, “A requirements taxonomy for reducing web site privacy vulnerabilities,” *Req’ts Engr. J.*, 9(3):169-185, 2004.

[2] A. Agresti, *Categorical Data Analysis*, Wiley, 2013.

[3] J. Bhatia, T.D. Breaux, F. Schaub. "Privacy goal mining through hybridized task re-composition", *ACM Trans. Soft. Engr. Method.*, 2016.

[4] J. Bhatia, M. Evans, S. Wadkar, T.D. Breaux “Automated extraction of regulated information types using hyponymy relations” *IEEE 3rd Int’l W‘shp on Artificial Intel. for Req’ts Engr.*, 2016.

[5] T.D. Breaux, H. Hibshi, A. Rao. "Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements." *Req’ts Engr. J.,* 19(3): 281-307, 2014.

[6] T.D. Breaux, A. Rao. “Formal Analysis of Privacy Requirements Specifications for Multi-Tier Applications,” *21st IEEE Int’l Req’ts Engr. Conf.,* pp. 14-23, Jul. 2013.

[7] T.D. Breaux, F. Schaub, “Scaling requirements extraction to the crowd: experiments on privacy policies,” *22nd IEEE Int’l Req’ts Engr. Conf.*, pp. 163-172, 2014.

[8] T.D. Breaux, D. Smullen, H. Hibshi. “Detecting Repurposing and Over-collection in Multi-Party Privacy Requirements Specifications.” *IEEE 23rd Int’l Req’ts Engr. Conf.*, pp. 166-175, Sep. 2015.

[9] L .F. Cranor, P. Guduru, and M. Arjula. 2006. “User interfaces for privacy agents,” *ACM Trans. Comput.-Hum. Interact.* 13, 2 (June 2006), pp. 135-178.

[10] L. Cranor, M. Langheinrich, M. Marchiori, M. Presler-Marshall, and J. Reagle. “The Platform for Privacy Preferences 1.0 (P3P1.0) Specification,” *World Wide Web Consortium Recommendation* April 2002. http://www.w3.org/TR/P3P

[11] M. C. Evans, J. Bhatia, S. Wadkar, T. D. Breaux. “An Evaluation of Constituency-based Hyponymy Extraction from Privacy Policies,” *Accepted to: 25th IEEE International Requirements Engineering Conference*, 2017.

[12] Q. He and A. I. Antón, “A Framework for Modeling Privacy Requirements in Role Engineering*,” International Workshop on Req’ts Engr. for Soft. Quality*, 16 - 17 June, 2003.

[13] A. M. McDonald and L. F. Cranor, “The cost of reading privacy policies”, *I/S – A Journal of Law and Policy for the Information Society, 4(3)*: 540-565, 2008.

[14] OECD, “The OECD Privacy Framework”, 2013.

[15] K. Pearson, “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” *Philosophical Magazine Series*, 5, 50 (302): 157–175, 1900.

[16] Andrew Perrin and Maeve Duggan, “Americans’ Internet Access: 2000-2015,” *Pew Research Center*, Washington, D.C., Date accessed: 22 June 2017. http://www.pewinternet.org/2015/06/26/americans-internet-access-2000-2015/

[17] J. Saldaña. *The Coding Manual for Qualitative Researchers*, SAGE Publications, 2012.

[18] M. C. Tschantz, A. Datta and J. M. Wing, "Formalizing and Enforcing Purpose Restrictions in Privacy Policies," *2012 IEEE Symposium on Security and Privacy*, San Francisco, CA, 2012, pp. 176-190.