Fer O'Neil, PhD student
Texas Tech University
Technical Writer, ESET

Analyzing Privacy Policies using the Privacy by Design Framework

Contents

Introduction

This project examines a select corpus of privacy policies using qualitative coding and data mining. Other research of privacy policies examines the language used from the perspective of *how* the information is communicated; that is, whether the information is readable (and typically whether it is machine readable). I chose the Privacy by Design (PbD) standard as a framework to analyze the content of privacy policies. The most important principle of PbD is "to keep it user-centric". That is, privacy policies exist to communicate how a person's information and data (from here I will use the term "information" for both) is collected, handled, and used by the companies that accumulate this information.

The results of this project will provide a first look at applying a PbD framework to analyze privacy policies and will suggest where to focus future research. Ultimately, by combining the results of a more comprehensive research using this approach with research on user preferences, we can advocate for changes to privacy policies. The changes will address not only how privacy policies should be communicated, but also how companies approach privacy with their products and services, how they

communicate this to their customers, and how we should consider personal information in the age of big data and The Internet of Things.

Research Goal

I examined the privacy policies of the top 10 most trusted companies for privacy to analyze whether the language they use to communicate their privacy policies conforms to the Foundational Principles of the Privacy by Design standard.

Methodology

Using data analysis software (QDA Miner), I assigned codes to each privacy policy and then analyzed the frequency of categories and codes. The three steps included the following:

1. Categorization and Coding
2. Frequency Analysis
3. Cluster and Correspondence Analysis

The initial coding used the sections and headings present within the policies to create the initial categories and codes. Subsequent passes through the policies yielded consistent categories and codes because most privacy policies contain similar sections, although they are not necessarily created from a uniform template. To ensure that the coding was applied consistently, codes were reviewed after the initial coding and recoded if inconsistencies were found. Because of the similarities among these cases, each policy coding virtually served as a recoding, and through the 12 cases, a consistent set of categories formed. Initially, I identified 11 categories and 30 codes. However, I further refined these when I applied the PbD Foundational Principles (see the next sub-section for a description for how I applied this coding) and the final coding identified 6 categories and 28 codes. Table 1 below displays the full list of final categories and codes.

Next, I used QDA Miner to examine the frequency of the categories and the number of occurrences of the codes. Coding for and subsequently analyzing the categories were used for the further analysis and interpretation of the categories and codes of interest (i.e., those that apply to the principles of Privacy by Design). The frequency analysis step was used to identify how often sections of interest were used in privacy policies.

Last, I performed an initial cluster analysis of the words that appeared within the most frequently occurring categories and codes. I say “initial” because this is an area that I need to examine more in future research (also see Appendix I). Cluster analysis examines the similarity of codes across policies, and by applying the PbD principles, it allows us to look more closely at the content within these identified codes.

Once words and phrases contained in the document were categorized and subjected to the consistency check, a frequency analysis was run at the Category level to determine how often policies referred to the coded topics. From the frequency analysis, I examined whether the most widely addressed category also contained the most words.

Data Sample

I chose to look at the privacy policies of "top 10 most trusted companies" because it is an established corpus (http://www.ponemon.org/blog/ponemon-institute-announces-results-of-2014-most-trusted-companies-for-privacy-study), and using an established corpus will allow me to focus on the analysis without projecting preconceived judgements regarding the "quality" of the content. Because of a tie, there are 12 policies and they are the following:

- Amazon www.amazon.com/privacy
- American Express https://www.americanexpress.com/us/content/legal-disclosures/online-privacy-statement.html
- PayPal https://www.paypal.com/us/webapps/mpp/ua/privacy-full
- Hewlett Packard http://www8.hp.com/us/en/privacy/privacy.html
- IBM http://www.ibm.com/privacy/details/us/en/
- Nationwide http://www.nationwide.com/privacy.jsp
- USAA https://www.usaa.com/inet/pages/privacy_promise?akredirect=true
- LinkedIn https://www.linkedin.com/legal/privacy-policy
- Apple http://www.apple.com/privacy/privacy-policy/?cid=wwa-us-kwg-features-com
- USPS https://about.usps.com/who-we-are/privacy-policy/welcome.htm
- Intuit https://security.intuit.com/privacy/
- Mozilla https://www.mozilla.org/en-US/privacy/

The web location of each policy was included in the Ponemon report and I accessed each one using a web browser. I then downloaded (if available from the website) or saved each policy to a Microsoft Word file as plain text. I then imported each file into QDA Miner.

Privacy by Design

I chose the PbD standard because its principles espouse similar user-centric goals to current technical communication theory. PbD was adopted as an international standard in a landmark resolution by the International Conference of Data Protection and Privacy Commissioners in Jerusalem ("Opinion on Privacy in the Digital Age: 'Privacy by Design' as a Key Tool to Ensure Citizens' Trust in ICTs." 2010).

Other research in this area has used standards such as The Platform for Privacy Preferences Project (P3P), which defines eight primary components with 11 purpose sub-elements to represent specific information use, combined with further attributes. The categories have similar categories and intentions to PbD (such as "how data is collected," which P3P labels "Purpose"). However, the project has been suspended in a "final state" since 2007.

PbD advances the view that we cannot assure the future of privacy solely by compliance with legislation and regulatory frameworks; rather, privacy assurance must become an organization's default mode of operation. The objectives of PbD — ensuring privacy and gaining personal control over one's information and, for organizations, gaining a sustainable competitive advantage — may be accomplished by practicing the 7 Foundational Principles ("7 Foundational Principles" 2015). However, I will not consider PbD from the product engineering perspective in this project and will instead focus on the communicative and end-user perspective because it is more than just an engineering guideline, and we must make this approach to "avoiding falling into techno-centric solutions to a socio-technical problem" (Gürses, Troncoso, and Diaz 2011, 5).

PbD is a blueprint, and Ann Cavoukian (the Information and Privacy Commissioner of Ontario who established the standard) leaves the foundational principles intentionally broad. With hope, this research will help identify some of the actual components of existing privacy policies and how they correlate to the PbD principles. PbD is user-centric, meaning it is contextual by nature, and we cannot merely create a compliance checklist that encompasses all policies, for all people. For this project, I have chosen four PbD principles to examine and have identified and assigned them codes from their broad descriptions. See Table 1 for final breakdown of categories, codes, and which PbD principles are identified to each code.

The four PbD principles discussed in this project and their descriptions are in the following table:

PbD Principle	Description of Principle
1) Proactive not Reactive; Preventative not Remedial (PbD-1)	The Privacy by Design approach is characterized by proactive rather than reactive measures. It anticipates and prevents privacy-invasive events before they happen.
2) Privacy as the Default Setting (PbD-2)	Privacy by Design seeks to deliver the maximum degree of privacy by ensuring that personal data are automatically protected in any given IT system or business practice.
6) Visibility and Transparency – Keep it Open (PbD-6)	Privacy by Design seeks to assure all stakeholders that whatever the business practice or technology involved, it is in fact, operating according to the stated promises and objectives, subject to independent verification.
7) Respect for User Privacy – Keep it User-Centric (PbD-7)	Privacy by Design requires architects and operators to keep the interests of the individual uppermost by offering such measures as strong privacy defaults, appropriate notice, and empowering user-friendly options.

Categories and Codes

Consistent themes emerged through the coding process and I assigned these to categories and codes. Most privacy policies are conventional, in that they all follow a consistent pattern and contain similar information, both anecdotally and confirmed by other research as well as this project (see Zimmeck and Bellovin 2014).

Table 1 lists each category, codes with the category, and which PbD principle I assigned. There were 11 codes that I identified from the four PbD principles examined in this project. Additionally, two of the PbD principles (PbD-2 and PbD-7) overlapped with three of the codes.

Table 1: Categories, codes, and related PbD principle

Category	Code	PbD Principle
What Personal Information is Collected		
	Automatic Information	PbD-1
	Mobile-Personal Info Collected	PbD-1
	Email Communications	
	Information You Give Us	
	Information from Third Party	PbD-1

	Affiliated Businesses	
	Cookies-General	
How Information is Used		
	Affiliated Businesses We Do Not Control	
	Third-Party Affiliated Service Providers	
	Promotional Offers	
	Normal Business Use	PbD-6
	For Our Protection-Comply with Law	
	With Your Consent	PbD-2
	Fraud	
	Selling-Disclosure	
How Information is Kept Secure		
	Data Security	
	Data Retention	
What Choices Do I Have		
	Not Provide Information	PbD-2, PbD-7
	Add or Update Certain Information	PbD-2, PbD-7
	Email or Mail Communication Preferences	PbD-7
	Not Allow Personal Information to Third-Party	PbD-2, PbD-7
	Cookies-Choices	PbD-7
	Opt-out	
	Access Information About Your Account	PbD-7
Minors		
	Children	
Privacy Complaints		
	Safe Harbor	
	File Complaint or Dispute	
	Truste	

Results and Discussion

After I coded and associated the PbD principles with the codes, I used QDA Miner to count the total occurrences of the identified PbD principles (see Table 2).

Table 2: Total count of PbD principles within categories

PbD Principle	Frequency of Category Occurrence (479)	% of Total
1) Proactive not Reactive; Preventative not Remedial (PbD-1)	62	12.90%
2) Privacy as the Default Setting (PbD-2)	30	6.00%
6) Visibility and Transparency – Keep it Open (PbD-6)	48	10.00%
7) Respect for User Privacy – Keep it User-Centric (PbD-7)	84	17.50%

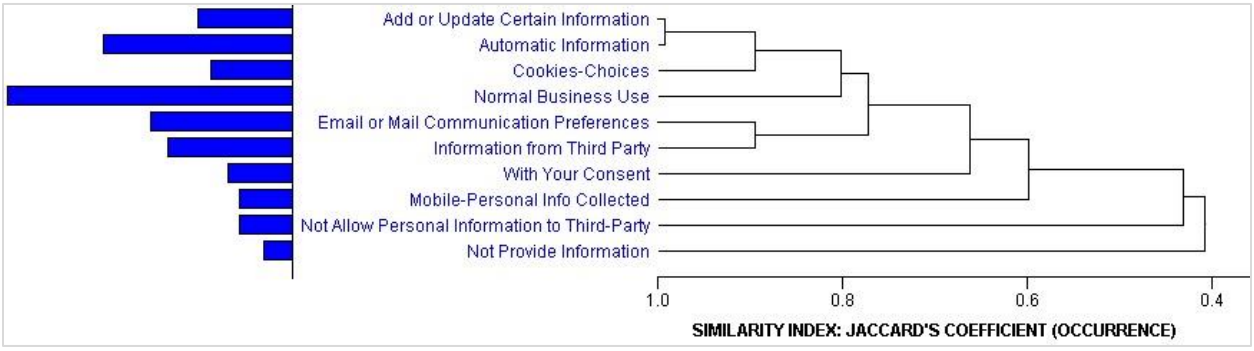
From the frequency count we can determine which codes are used more often, and more importantly, when and whether the PbD-identified codes are used. Table 3 shows a list of the most frequently used codes, with the PbD principles identified with each.

Table 3: Count and percent occurrence of codes to total cases (12)

Category	Code	Description	Count	Cases	% Cases
How Information is Used	Normal Business Use	PbD-6	48	12	100.00%
What Personal Information is Collected	Automatic Information	PbD-1	32	10	83.30%
What Personal Information is Collected	Information from Third Party	PbD-1	21	10	83.30%
What Choices Do I Have	Add or Update Certain Information	PbD-2, PbD-7	16	10	83.30%
What Choices Do I Have	Email or Mail Communication Preferences	PbD-7	24	9	75.00%
What Choices Do I Have	Cookies-Choices	PbD-7	14	9	75.00%
How Information is Used	With Your Consent	PbD-7	11	8	66.70%
What Personal Information is Collected	Mobile-Personal Info Collected	PbD-1	9	6	50.00%
What Choices Do I Have	Not Allow Personal Information to Third-Party	PbD-2, PbD-7	9	5	41.70%
What Choices Do I Have	Not Provide Information	PbD-2, PbD-7	5	4	33.30%
What Choices Do I Have	Access Information About Your Account	PbD-7	5	3	25.00%

Looking at the data, we can see the distribution of PbD principles across the cases, although we cannot argue whether this contributes to policies that lack the user-centric goal of the PbD standard. As Table 3 shows, the most commonly used code is “Normal Business Use” and this is identified as PdB-6. This code is the highest occurring (frequency, see Table 3 above) and is second only to Automatic Information in number and percentage of words (also see Figure 3 below). The Normal Business Use code is closely associated with Cookies-Choices (PbD-7), Automatic Information (PbD-1), and Add or Update Certain Information (PbD-2, PbD-7) and seems to be a catch-all category for any reason for using personal data that does not apply to another code (see Figure 1 below).

Figure 1: Similarity between and among codes



However, we cannot determine based on this data an overall positive or negative aspect to this code's use. The results do, however, indicate that this is an element of privacy policies that needs further investigation. For example, we can identify areas of interest within this corpus of privacy policies by coding for and subsequently analyzing the categories to determine codes of interest (i.e., those that apply to the principles of Privacy by Design, see Table 2 above) and whether these codes are associated with PbD principles.

Next, we can use a frequency analysis to investigate whether the most widely addressed category will include, for example, the highest word count (see Figure 3 below). This analysis is important to determine if a high frequency category is the most fundamental element of a privacy policy. Also of note will be those categories that are not used often, and to interpret why (e.g., are they user-focused or business focused).

PbD advocates for keeping the interests of the individual central "by offering such measures as strong privacy defaults, appropriate notice, and empowering user-friendly options" (Pagallo 2012, 16). The following four codes are those that we can identify as conforming to the PbD principles for user-centric policies:

- Add or Update Certain Information (PbD-2, PbD-7)
- Email or Mail Communication Preferences (PbD-7)
- Not Allow Personal Information to Third-Party (PbD-2, PbD-7)
- Cookies-Choices (PbD-7)

PbD-7 may be the most applicable for performing a future content analysis on the specific categories and codes that apply to users. In this case (see Figure 3 below), the highest percentage of words are within non-user-centric codes (Cookies-General). Cookies-Choices is the first user-centric principle, but this is the de-facto "token" user-centric principle in most privacy policies (see Figure 2). Opt-out is not user-centric because PbD states that everything should be opt-in

Figure 2: Example of opt-out of cookies on website

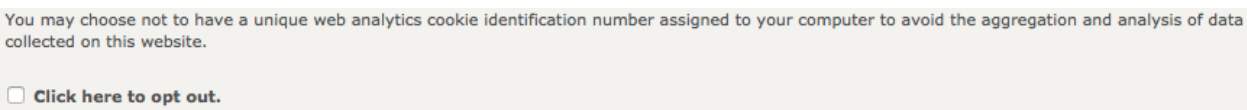
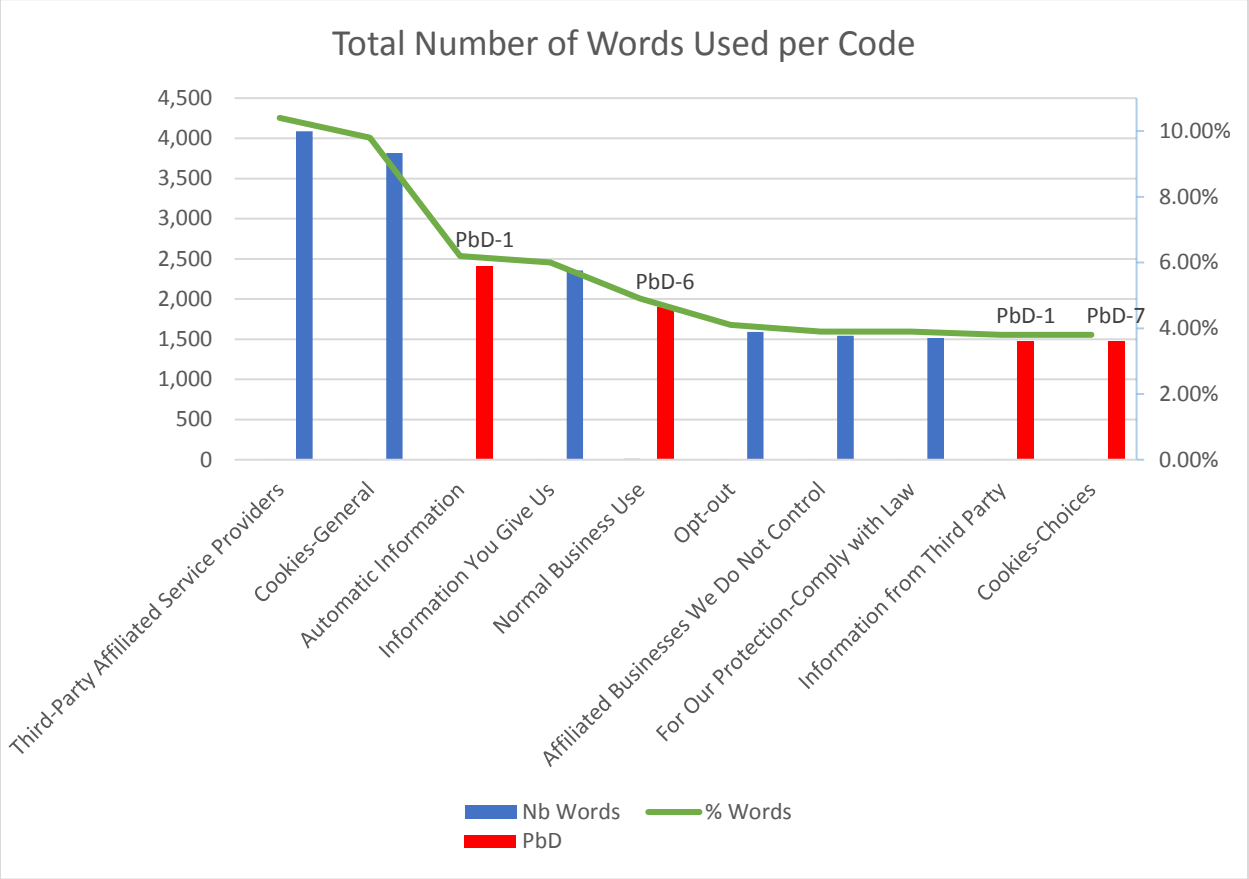


Figure 3: Top codes with the total number of words used in each (and percentage of the code to the total) with the PbD codes identified in red



As Figure 3 displays, there are more words used in the business-focused codes (blue), which could indicate a preference for obtaining the data associated with the code, or a compliance reason for communicating about it. From this data, we cannot correlate less communication of the PbD codes to less user-centric policies, but I think that these results warrant a closer examination. Future research will include a content analysis of these selected codes to help make such a determination.

Conclusion, Limitations, and Future Research

If a privacy policy does not explicitly discuss certain information, it is possible that that information is covered by existing laws, rules, or regulations (i.e., ignorance of the law is no excuse) and, therefore, we cannot determine with certainty that something missing or lacking within a policy is not covered somewhere. In addition to acknowledging what is missing in policies, determining the most common categories to code within them will aid further research in this area. A full genre analysis of privacy policies would contribute to a definitive list of categories. Additionally, a comprehensive literature review of the research on privacy policies will help to frame and guide this research. For example, from the literature we can see what information is most important to users (see "Know Privacy" 2015 and section "Complex Privacy Preferences" in Cranor, Guduru, and Arjula 2006, 142-143) and when combined with the analysis performed in this project for what information companies communicate, the most common elements are 1) what information is collected 2) how that information

is used. Therefore, with such a robust study that include the additional components above (i.e., genre analysis, literature review, combined with text mining research), we can begin to make claims regarding the efficacy of what *is included* in existing privacy policies.

This pilot project does contribute to a better understanding of what information is included in the privacy policies of the most trusted companies (Ponemon Institute 2015) because the results do suggest that the percent of codes closely correlates to the number of and percent of total cases that the code appears in. That is, the more times a coding category appears in total throughout all the policies, the more likely it is to appear in all policies. Proceeding with the preceding intimation, when combined with the PbD framework we can start to make recommendations for what is included, and what is missing as well.

There are additional areas of interest that arose from this project that could be researched further. For example, the "data retention" and "selling-disclosure" content was interesting as it pertains to current events. Recently, RadioShack filed for bankruptcy and attempted to sell the personally-identified data collected from its customers over its 90 year history[1]. I foresee this issue I identified within this corpus of privacy policies to continue to make news as current law and society's perceptions attempt to come to terms with how personally identifiable information is controlled. The "Know Privacy" project addressed the lack of dependable information in privacy policies concerning collected information:

> Additionally, very few of them made clear statements about the fate of user data in the event of a merger or bankruptcy, or if they enhance the data by purchasing information about users from outside sources to build more detailed profiles" ("Know Privacy" 2015, Privacy Policies).

However, it seems that people are not concerned enough at the moment for a paradigm change to an opt-in culture of privacy policies. Further research on this topic would include an analysis of PbD-4, "Full Functionality – Positive-Sum, not Zero-Sum." Last, there is a lack of current research on user preferences for what privacy issues are important to them and how they want to control their preferences on websites. In order to recommend changes to privacy policies to be user-centric, we would need to combine not only the PbD framework (because it is a standard that does not take user preference into account in each, individual instance) but also the user testing and feedback.

References

"7 Foundational Principles." 2015. *Privacy By Design*. Accessed August 4. https://www.privacybydesign.ca/index.php/about-pbd/7-foundational-principles/.

Cranor, Lorrie Faith, Praveen Guduru, and Manjula Arjula. 2006. "User Interfaces for Privacy Agents." *ACM Trans. Comput.-Hum. Interact.* 13 (2): 135–78. doi:10.1145/1165734.1165735.

Gürses, Seda, Carmela Troncoso, and Claudia Diaz. 2011. "Engineering Privacy by Design." *Computers, Privacy & Data Protection* 14.

"Know Privacy." 2015. Accessed August 2. http://knowprivacy.org/.

[1] http://www.theguardian.com/business/2015/may/18/bankrupt-radioshack-selling-customer-data-names-addresses

"Opinion on Privacy in the Digital Age: 'Privacy by Design' as a Key Tool to Ensure Citizens' Trust in ICTs." 2010. European Data Protection Supervisor. https://secure.edps.europa.eu/EDPSWEB/edps/EDPS/Pressnews/News/N2010.

Pagallo, Ugo. 2012. "On the Principle of Privacy by Design and Its Limits: Technology, Ethics and the Rule of Law." In *European Data Protection: In Good Health?*, edited by Serge Gutwirth, Ronald Leenes, Paul De Hert, and Yves Poullet, 331–46. Springer Netherlands. http://link.springer.com/chapter/10.1007/978-94-007-2903-2_16.

Ponemon Institute. 2015. "Ponemon Institute Examines Consumer Response to Data Breach Notice - News and Press Releases." Accessed May 25. http://www.ponemon.org/news-2/4.

Zimmeck, Sebastian, and Steven M. Bellovin. 2014. "Privee: An Architecture for Automatically Analyzing Web Privacy Policies." In *23rd USENIX Security Symposium (USENIX Security 14)*, 1–16. San Diego, CA: USENIX Association. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/zimmeck.