Visualization and Interactive Exploration of Data Practices in Privacy Policies

Sushain K. Cherivirala
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
sushain@cs.cmu.edu

Shomir Wilson
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
shomir@cs.cmu.edu

Florian Schaub
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
fschaub@cs.cmu.edu

Norman Sadeh
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
sadeh@cs.cmu.edu

Mads Schaarup Andersen
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
manderse@cs.cmu.edu

Joel R. Reidenberg
School of Law
Fordham University
New York, NY, USA
jreidenberg@fordham.edu

Poster presented at the 12th Symposium on Usable Privacy and Security (SOUPS 2016), June 22-24, 2016, Denver CO.

Abstract
The Usable Privacy Policy Project researches methods and techniques to semi-automatically analyze natural language privacy policies and extract data practices described in them. This effort aims to investigate and improve the effectiveness of notice and choice by informing public policy and providing data practice information in more usable notice formats to users. We present explore.usableprivacy.org – a website that visualizes data practices in privacy policies and facilitates the interactive exploration of a privacy policy's content. Our website is based on a corpus of 115 annotated privacy policies, which cover 192 websites, mobile apps, and service providers. Our website constitutes a valuable tool for privacy researchers, activists, and regulators to gain insights on the structure, composition, and content of privacy policies.

Author Keywords
Privacy; privacy policy; usability; visualization.

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Introduction
Natural language privacy policies have become a de facto standard to provide "notice and choice" to users and con-

sumers. Yet, there is ample evidence that users generally do not read these policies and that those who occasionally do struggle to understand what they read [2, 7, 3, 8, 10]. Previous efforts on making the information in privacy policies more accessible focused on machine-readable formats, such as the Platform for Privacy Preferences (P3P) [11], or on improving the usability of notice formats [5, 6, 10]. However, such efforts rely on industry action but often lack adoption incentives.

The Usable Privacy Policy Project[1] investigates methods and techniques to analyze natural language privacy policies by combining expert annotations and crowdsourcing with natural language processing and machine learning [4, 13, 9, 1]. The goal is to semi-automatically extract relevant data practices described in the privacy policies, and present those features to users in an easy-to-digest format that enables them to make more informed privacy decisions as they interact with websites, mobile apps, and other services. As part of this project, we compiled a corpus of 115 privacy policies with a total of 23,000 data practice statements annotated by experts [12]. We created a website – explore.usableprivacy.org – that visualizes these annotated data practices in the context of the original privacy policies and facilitates the interactive exploration of this rich dataset (see Figure 1). The website provides the opportunity to gain insights on the structure, composition, and content of privacy policies – for specific websites and across website categories. This makes it a valuable tool for interested users, privacy researchers, activists, and regulators.

Next, we shortly introduce our corpus of annotated privacy policies, before describing in more detail how our explore website visualizes policy content and facilitates its exploration.

[1] UPPP website: www.usableprivacy.org

Figure 1: Policy exploration website (explore.usableprivacy.org).

Privacy Policy Annotation Corpus

The data shown on our privacy policy exploration website is the result of a large-scale annotation effort of privacy policies from U.S.-based websites [12]. We developed a frame-based annotation scheme to represent and extract privacy practice statements from privacy policies. We capture nine categories of data practices: first part collection/use; third party sharing/collection; user choice/control; user access, edit and deletion; data retention; data security policy change; Do Not Track; and provisions for international and specific audiences (e.g., children or California residents). Each data practice category consists of a set of practice attributes (e.g., information type, purpose), and specific values (e.g., contact information).

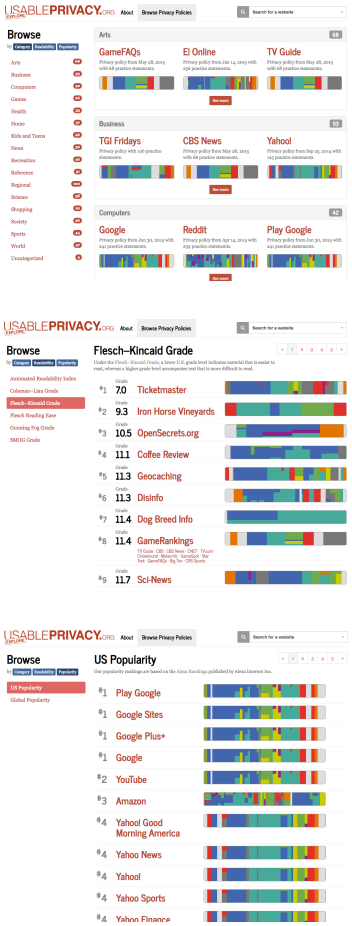

Figure 2: Privacy policies and websites can be browsed by website category (*top*), readability scores (*middle*), and popularity ranking (*bottom*).

We collected a diverse set of 115 privacy policies, covering 192 websites. These websites stem from a diverse set of website categories (all of DMOZ.org's top-level categories are represented) and include highly popular as well as long-tail websites (according to their Alexa.com rank). Each of these privacy policies was annotated by three experts. We hired ten law students as expert annotators. Our annotators used an online annotation tool, which presented policies paragraph by paragraph and asked them to mark all described data practices, by selecting respective categories and selecting the attributes that describe the practice. For each attribute value, annotators also marked the respective policy text.

In total, our corpus of 115 policies contains 23K annotated data practice statements, 128K annotated practice attributes, and 103K annotated text spans. Note that each policy was annotated by three annotators. This rich corpus of annotations can serve as an evaluation standard for research on machine learning, natural language processing and crowdsourcing. It is also intended to provide insights into the composition and content of privacy policies.

Privacy Policy Exploration Website

Our privacy policy exploration website currently has two main functionalities: (1) facilitate navigation through the corpus of annotated privacy policies and the covered websites, and (2) visualization and in-depth exploration of a specific privacy policy's content. Thus, the site's homepage, shown in Figure 1, prominently displays a website search bar and three randomly-selected example privacy policies, including a miniature representation of a policy's composition – colors indicate where statements from different data practice categories can be found in the privacy policy.

Our website enables users to browse annotated privacy policies and websites based on website category, readability scores, and popularity ranking (see Figure 2). When navigating to the browse page, the category view is shown by default. It lists all website categories (retrieved from DMOZ.org) and provides an overview of the websites analyzed in each category. By grouping websites into categories, we facilitate the comparison of similar websites. Note that a website may appear in multiple categories. The other two browse views facilitate surveying websites and their privacy policies ordered by either their readability score (e.g., Flesch-Kincaid) or their US or global popularity rank (according to Alexa.com)

Selecting a specific website opens the respective details page. These details pages are the primary sources of information on the policy exploration website. Figure 3 shows the details page for The New Yorker's privacy policy.

The details page consists of four main parts. The website name, URL and the website's categories are listed at the top. The left column provides access to the annotated privacy practices. The center part shows the actual privacy policy text. The policy's metadata such as its collection or last updated date, readability score and practice statement count is provided above the policy's text. While we display the Flesch-Kincaid Grade Level by default, clicking on the grade level opens a modal dialog with further readability scores. The policy metadata further includes the company's name, because multiple websites may be covered by the same privacy policy. In this example, newyorker.com is covered by the Conde Nast privacy policy; other websites covered by the same policy are also listed as part of the policy metadata. Annotated data practices are marked with color highlights in the privacy policy text. A minimap on the right of the policy text provides an overview of the policy's com-

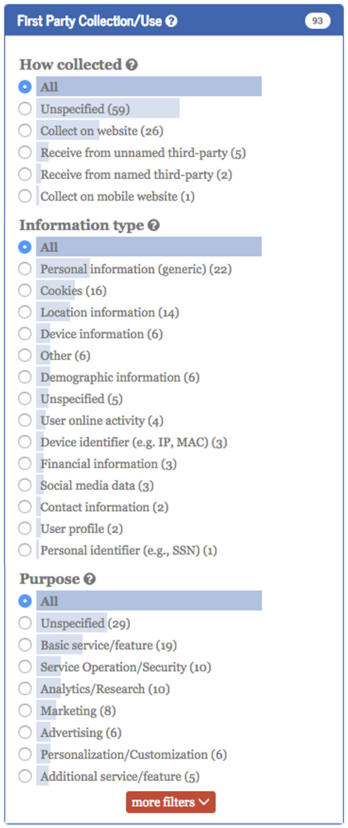

Figure 4: Expanding a data practice category reveals category-specific attribute statistics and filtering options.

Figure 3: Privacy policy details page for The New Yorker.

position and provides navigational support when scrolling. The policy minimap can also be clicked to jump to a specific position. To ease first time use, we show an interactive tour of the interface on a user's first visit to explain the multiple interactive components.

The privacy practices overview to the left of the policy text facilitates deeper investigation of the policy content. Aside from visualizing the frequency of statements in each category, individual categories can be expanded to reveal detailed statistics and filtering options for the category-specific practice attributes and attribute values. Figure 4 shows the first party collection/use category expanded. These filtering options allow users to refine which data practice statements are highlighted in the policy text and the minimap. Clicking on any practice statement highlighted in the policy text expands the corresponding practice category. It also displays a textual summary of the data practice, constructed using a template-based approach for natural language generation, to help readers interpret the policy text.

With the tools provided by the policy details page, we believe users are equipped to quickly locate specific information within a privacy policy, and drill-down to specific practice expressions, for example, statements that indicate collection of information type, but are not clear on the purpose or how it is collected ("Unspecified").

Conclusions & Future Work

With the privacy policy exploration website, we make a rich corpus of annotated privacy policies easily accessible to privacy researchers, activists, regulators, and the general public. Websites covered by annotated privacy policies can be scrutinized for egregious practices, vague or abstract descriptions, as well as privacy-friendly practices.

We plan to further improve our website by integrating semantic analysis and search functionalities, which would allow to compose more complex queries and retrieve respective annotated data practices from multiple privacy policies. In general, we plan to ease comparison of data practices across multiple websites, as this also holds the potential to highlight similar, yet privacy-friendlier websites.

Making raw annotation data available for download is also a priority to facilitate further analysis and research with our corpus. At the same time, we are also interested in integrating features that enable assessment of policies by the community and regular users. Our hope that interactive exploration tools can help stipulate and increase engagement with privacy policies – documents everyone implicitly consents to everyday when using any service, most without reading them.

Acknowledgements

This work has been funded by the National Science Foundation under grants CNS-1330596 and CNS-1330214. The authors would like to acknowledge all members of the Usable Privacy Policy Project (www.usableprivacy.org) for their contributions.

REFERENCES

1. Travis. D. Breaux and Florian Schaub. 2014. Scaling requirements extraction to the crowd: Experiments with privacy policies. In *Int. Req. Eng. Conf. (RE'14)*. IEEE.
2. Fred H. Cate. 2010. The Limits of Notice and Choice. *IEEE Security & Privacy* 8, 2 (March 2010), 59–62.
3. Lorrie Faith Cranor. 2005. Giving notice: Why privacy policies and security breach notifications aren't enough. *IEEE Communications Magazine* 43, 8 (Aug. 2005), 18–19.
4. Norman Sadeh et al. 2013. *The Usable Privacy Policy Project: Combining Crowdsourcing, Machine Learning and Natural Language Processing to Semi-Automatically Answer Those Privacy Questions Users Care About*. Tech. report CMU-ISR-13-119. Carnegie Mellon University.
5. Loretta Garrison, Manoj Hastak, Jeanne M. Hogarth, Susan Kleimann, and Alan S. Levy. 2012. Designing Evidence-based Disclosures: A Case Study of Financial Privacy Notices. *Journal of Consumer Affairs* 46, 2 (June 2012), 204–234.
6. Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. 2010. Standardizing privacy notices: an online study of the nutrition label approach. In *Proc. CHI '10*. ACM.
7. Aleecia M. McDonald and Lorrie Faith Cranor. 2008. The Cost of Reading Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society* 4, 3 (2008), 540–565.
8. President's Concil of Advisors on Science and Technology. 2014. *Big Data and Privacy: A Technological Perspective*. Report to the President. Executive Office of the President.
9. Joel R. Reidenberg, N. Cameron Russell, Alexander J. Callen, Sophia Qasir, and Thomas B. Norton. 2015. Privacy Harms and the Effectiveness of the Notice and Choice Framework. *I/S Journal of Law & Policy for the Information Society* 11, 2 (2015).
10. Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor. 2015. A Design Space for Effective Privacy Notices. In *Proc. Symp. Usable Privacy and Security (SOUPS'15)*.
11. Rigo Wenning, Matthias Schunter, Lorrie Cranor, B. Dobbs, S. Egelman, G. Hogben, J. Humphrey, M. Langheinrich, M. Marchiori, M. Presler-Marshall, J. Reagle, and D. A. Stampley. 2006. The Platform for Privacy Preferences 1.1 (P3P 1.1) Specification. (2006). `http://www.w3.org/TR/P3P11/`
12. Shomir Wilson, Florian Schaub, Aswarth Dara, Sushain K. Cherivirala, Sebastian Zimmeck, Mads Schaarup Andersen, Pedro Giovanni Leon, Eduard Hovy, and Norman Sadeh. 2016a. Demystifying Privacy Policies Using Language Technologies: Progress and Challenges. In *TA-COS '16: LREC Workshop on Text Analytics for Cybersecurity and Online Safety*.
13. S. Wilson, F. Schaub, R. Ramanath, N. Sadeh, F. Liu, N.A. Smith, and F. Liu. 2016b. Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?. In *Int. World Wide Web Conference (WWW '16)*.