

The Creation and Analysis of a Website Privacy Policy Corpus

Shomir Wilson¹, Florian Schaub¹, Aswarth A. Dara¹, Frederick Liu¹, Sushain Chervirala¹, Pedro G. Leon², Mads S. Andersen¹, Sebastian Zimmeck³, Kanthashree M. Sathyendra¹, N. Cameron Russell⁴, Thomas B. Norton⁴, Eduard Hovy¹, Joel R. Reidenberg⁴, Norman Sadeh¹

¹Carnegie Mellon University, School of Computer Science, Pittsburgh, PA

²Stanford University, Center for Internet and Society, Stanford, CA

³Columbia University, Department of Computer Science, New York, NY

⁴Fordham University, Law School, New York, NY

Extended Abstract

Website privacy policies are often ignored by Internet users, because these documents tend to be long and difficult to understand. However, the significance of privacy policies greatly exceeds the attention paid to them: these documents are binding legal agreements between website operators and their users, and their opaqueness is a challenge not only to Internet users but also to policy regulators. One proposed alternative to the status quo is to automate or semi-automate the extraction of salient details from privacy policy text, using a combination of crowdsourcing, natural language processing, and machine learning. However, there has been a relative dearth of datasets appropriate for identifying data practices in privacy policies.

To remedy this problem, we created a corpus of 115 privacy policies (267K words) with fine-grained manual annotations of data practices by law students. Our dataset contains 23,000 fine-grained data practice annotations for ten different types of data practices. Each data practice consists of multiple attributes (e.g., information collected, purposes of collection, collection context, etc.) and each attribute value is accompanied by supporting policy text selected by the annotator. The website <https://explore.usableprivacy.org> makes this dataset available to the public and the research community. The website allows to browse and visually explore the annotated privacy policies with their data practice annotations.

We describe the process of using skilled annotators and a purpose-built annotation tool to produce the data. We provide findings based on a census of the annotations and show results toward automating the annotation procedure. Finally, we describe challenges and opportunities for the research community to use this corpus to advance research in both privacy and language technologies.

This research was first published in the *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL '16)* – August 2016.

Full paper available at: <http://www.aclweb.org/anthology/P/P16/P16-1126.pdf>