The PrivaSeer Project: Large-Scale Resources for Analysis of Privacy Policy Text

Shomir Wilson
Pennsylvania State University

Florian Schaub
University of Michigan

Lee Matheson
Future of Privacy Forum

Shahriar Shayesteh
Pennsylvania State University

Lu Xian
University of Michigan

Abstract

Privacy policies provide insight into organizations' data processing practices, but the wealth of privacy policies available on the web contrasts with the challenges of understanding the state of digital privacy at scale. We report on progress made by the PrivaSeer Project (https://privaseer.ist.psu.edu/) to build large-scale, longitudinal, annotated, and usable resources for the study of website privacy policies. These resources are aimed at privacy researchers, practitioners, and policymakers, a set of groups with varying technical backgrounds and analysis goals. We describe the *PrivaSeer Corpus*, the largest to-date publicly available corpus of privacy policies, and *PrivaSeer Search*, a search engine that makes browsing and exploring the corpus easy for a variety of stakeholders. We also summarize analysis of privacy policy availability, languages privacy policies are written in, and the prevalence of dates in privacy policies. These results provide a large-scale snapshot of the contents of privacy policies, with implications for their usability and legal compliance.

1 Motivation and Overview

A continuing obstacle to technology users' understanding of digital privacy is the privacy policies they encounter. These documents, written in natural language, describe the data collection practices of an organization and the choices that users can make about those practices. Privacy policies in their typical form are a challenge to users' understanding, but they are also a challenge to work by privacy researchers, policymakers, and practitioners [6]. All these groups have a stake in making digital privacy more usable, moving beyond a long-standing status quo of internet users struggling to understand privacy policies [4, 5] and typically not reading them [8].

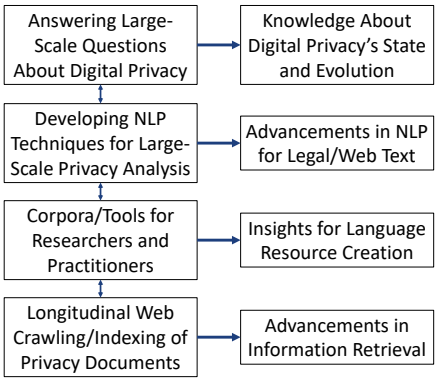

Figure 1: Overview of research performed by the PrivaSeer Project (left items) and broader impacts (right).

We present recent work by the PrivaSeer Project to create large-scale, longitudinal, annotated, and usable resources for studying websites' privacy policies written in English. Figure 1 provides an overview of the project's research and broader impacts. These cover a multidisciplinary range of activities in information retrieval, natural language processing (NLP), and public policy. Although prior privacy policy collections exist [2, 15, 16], PrivaSeer resources cover a substantially larger number of unique websites. We describe the project's work to gather privacy policies into the PrivaSeer Corpus, to index privacy policies in an easy-to-use search engine (PrivaSeer Search), and to perform large-scale studies of privacy policies across the web.

2 Resources

The PrivaSeer Corpus is the basis for the project's work. To create the corpus, the team gathered likely privacy policy URLs from the Common Crawl dataset [3] based upon

USENIX Symposium on Usable Privacy and Security (SOUPS) 2025.
August 10–12, 2025, Seattle, WA, United States.

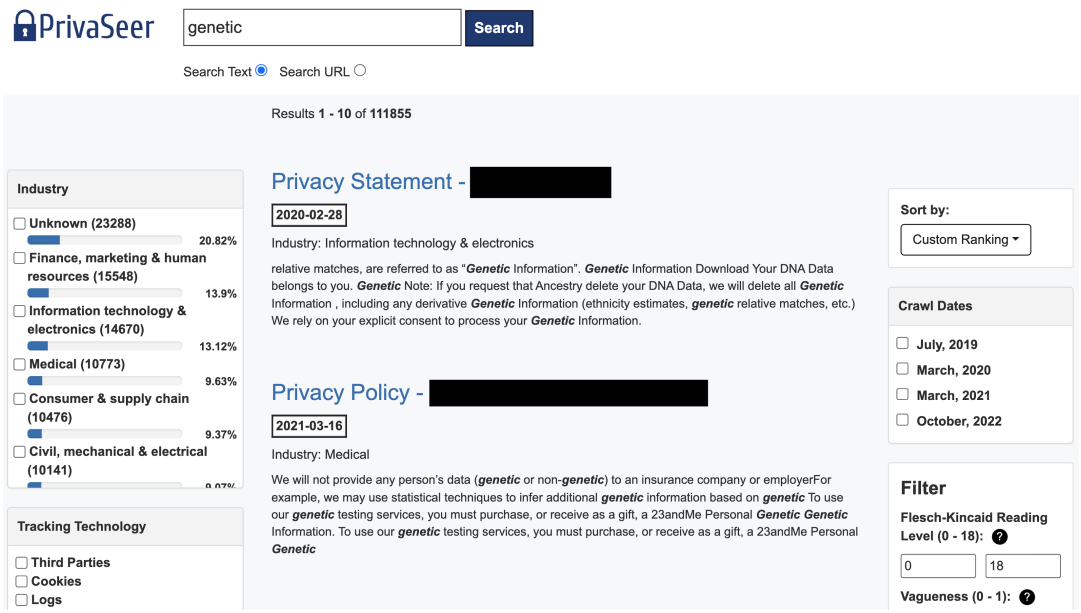

Figure 2: A cropped screen shot of the PrivaSeer Search interface. Websites' names are redacted from the image.

keyword heuristics. The team created an automated pipeline to download web pages at each URL, filter out non-English pages[1], perform text classification to eliminate documents unlikely to be privacy policies, and eliminate duplicates. The same pipeline was used for subsequent updates, and as of April 2025, The PrivaSeer Corpus contains 3,967,487 privacy policies. The corpus is publicly available on the project website [9] under a Creative Commons BY-NC-SA license. More details, including performance statistics for heuristics in the collection pipeline, are available in a dedicated paper [13].

Figure 1 shows PrivaSeer Search, a search engine that indexes privacy policies collected by the project. Through a variety of search facets, users can filter results by industry, crawl date, readability metrics, and mentions to tracking technologies, self-regulatory bodies, and laws. The search engine allows researchers with limited technical knowledge to interact with the corpus and enables others to explore the corpus before downloading it. The search engine is publicly available [9], and a dedicated paper [12] describes its creation and evaluation.

3 Analysis

The team used the PrivaSeer infrastructure to study privacy policy availability across 7M domains on the web [11]. A similar corpus creation pipeline led to the estimate that 34% of web domains link to a privacy policy on their landing pages. Privacy policy availability varied by sector of activity, with medical websites most likely to post a privacy policy. Also, mismatches were observed between the natural languages of websites' landing pages and their privacy policies. A small number of purported privacy policies were in Latin, which is unsuitable for legal notices in any modern jurisdiction, though most consisted of *lorem ipsum* placeholder text [14].

Another study focused on dates in privacy policy text [10], driven by the observation that many privacy laws (including GDPR [1]) require privacy policies to remain up-to-date. An examination of 3.5M privacy policies showed that only 40% contained one or more dates. The chronological distribution of dates showed a substantial increase in 2017-2019, likely due to GDPR's inception in 2018. Additionally, websites with higher PageRank [7] were more likely to contain at least one date, showing a correlation between a website's popularity and attendance to legal obligations.

Finally, an analysis of privacy policies in the U.S. financial sector found fragmented disclosures. Almost half of the largest banks provided multiple privacy documents, and many contradict themselves across disclosures, often between the GLBA notice and CCPA privacy policy. These inconsistencies, shaped by overlapping privacy regulations, risk misleading consumers and show a need for standardized disclosure requirements.

Acknowledgments

This research was supported in part by grants from the National Science Foundation's Secure and Trustworthy Computing program (CNS-2105745, CNS-2105734, CNS-2105736).

[1]The purpose of the English constraint was to match the language expertise of the researchers. However, many non-English privacy policies were also collected, and we welcome collaborations to study them.

References

[1] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016. Accessed: 2025-04-20.

[2] Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. In *Proceedings of The Web Conference 2021*, WWW '21, page 22. Association for Computing Machinery.

[3] Common Crawl. Common crawl dataset. https://commoncrawl.org, 2024. Accessed: 2025-04-20.

[4] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. Large-scale readability analysis of privacy policies. In *Proceedings of the International Conference on Web Intelligence*, WI '17, page 18–25, New York, NY, USA, 2017. Association for Computing Machinery.

[5] Aleecia McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 4:543, 2008.

[6] Abraham Mhaidli, Selin Fidan, An Doan, Gina Herakovic, Mukund Srinath, Lee Matheson, Shomir Wilson, and Florian Schaub. Researchers' experiences in analyzing privacy policies: Challenges and opportunities. *Proceedings on Privacy Enhancing Technologies*, 2023:287–305, 10 2023.

[7] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford InfoLab, 1999.

[8] President's Council of Advisors on Science and Technology. Big data and privacy: A technological perspective. Report to the president, Executive Office of the President, May 2014.

[9] PrivaSeer. PrivaSeer. https://privaseer.ist.psu.edu/, 2025. Accessed: 2025-04-20.

[10] Mukund Srinath, Lee Matheson, Pranav Narayanan Venkit, Gabriela Zanfir-Fortuna, Florian Schaub, C. Lee Giles, and Shomir Wilson. Privacy now or never: Large-scale extraction and analysis of dates in privacy policy text. In *Proceedings of the ACM Symposium on Document Engineering 2023*, DocEng '23, New York, NY, USA, 2023. Association for Computing Machinery.

[11] Mukund Srinath, Soundarya Sundareswara, Pranav Venkit, C. Lee Giles, and Shomir Wilson. Privacy lost and found: An investigation at scale of web privacy policy availability. In *Proceedings of the ACM Symposium on Document Engineering 2023*, DocEng '23, New York, NY, USA, 2023. Association for Computing Machinery.

[12] Mukund Srinath, Soundarya Nurani Sundareswara, C. Lee Giles, and Shomir Wilson. Privaseer: A privacy policy search engine. In Marco Brambilla, Richard Chbeir, Flavius Frasincar, and Ioana Manolescu, editors, *Web Engineering - 21st International Conference, ICWE 2021, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 286–301, Germany, 2021. Springer Science and Business Media Deutschland GmbH. Publisher Copyright: © 2021, Springer Nature Switzerland AG.; 21st International Conference on Web Engineering, ICWE 2021 ; Conference date: 18-05-2021 Through 21-05-2021.

[13] Mukund Srinath, Shomir Wilson, and C Lee Giles. Privacy at scale: Introducing the PrivaSeer corpus of web privacy policies. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6829–6839, Online, August 2021. Association for Computational Linguistics.

[14] Wikipedia contributors. Lorem ipsum. https://en.wikipedia.org/wiki/Lorem_ipsum, 2025. Accessed: 2025-04-20.

[15] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. The creation and analysis of a website privacy policy corpus. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, Berlin, Germany, August 2016. Association for Computational Linguistics.

[16] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R. Reidenberg, N. Cameron Russell, and Norman Sadeh. Maps: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies*, 2019(3):66–86, 2019.