

Privacy Later: Reflections on Design Decisions, Lessons Learned, and Open Problems of the OPP-115 Corpus a Decade After Release

Shomir Wilson¹, Florian Schaub², Norman Sadeh³

¹Pennsylvania State University, University Park, PA, USA, shomir@psu.edu

²University of Michigan, Ann Arbor, MI, USA, fschaub@umich.edu

³Carnegie Mellon University, Pittsburgh, PA, USA, sadeh@cs.cmu.edu

Abstract

The OPP-115 Corpus of annotated privacy policies was released by the Usable Privacy Policy Project in 2016 and was the first large-scale effort to annotate privacy policies in a fine level of detail. Since then, it has been used by natural language processing (NLP) researchers working on automated analysis of websites’ and applications’ privacy policy documents, with efforts continuing in the present day. We share reflections on design decisions that went into the creation of the corpus and lessons learned from obstacles associated with it. Additionally, we describe open problems in research on privacy documents that the OPP-115 Corpus introduced or brought attention toward. These open problems show the continuing need for research on privacy policy text to support consumers’ digital privacy.

1 Introduction

Websites’ and apps’ privacy policies pose a variety of obstacles to interpretation. For consumers, they are lengthy, difficult and time-consuming to read (McDonald and Cranor, 2008; Reidenberg et al., 2015; Degeling et al., 2019), hindering effective notice and choice (Cranor, 2012). For law scholars and privacy regulators, they are vague and also challenging to interpret at scale (Reidenberg et al., 2016), hindering research into the state and evolution of digital privacy practices (Mhaidli et al., 2023). This confluence has motivated a steady stream of research in natural language processing (NLP), privacy, and human-computer interaction to scalably and accurately interpret privacy policies and present them to stakeholders in ways that respond to their needs (Ravichander et al., 2021).

An early progenitor of NLP research on privacy policies was the OPP-115 Corpus (Online Privacy Policies, set of 115), which consisted of 115 website privacy policies each manually annotated by three law students (Wilson et al., 2016b). This cor-

pus was released in 2016 by the US NSF-funded Usable Privacy Policy Project¹ as part of a series of efforts to study manual, semi-automated, and fully automated information extraction from privacy policies (Sadeh et al., 2013). The corpus’ frame-based annotation scheme yielded 23K *data practices* (i.e., representations of how a data controller collects, uses, or shares personally identifiable information), which in aggregate yielded 128K attribute selections and 103K text spans associated with those selections.

Figure 1 shows annual counts of citations to the ACL 2016 paper that introduced the corpus (Wilson et al., 2016b), totaling 540 in Google Scholar² as of September 19, 2026. The corpus not only enabled NLP research on privacy policies but also inspired additional corpora, such as PrivacyQA (Ravichander et al., 2019), PolicyQA (Ahmad et al., 2020), the Princeton-Leuven longitudinal corpus (Amos et al., 2021), and PrivaSeer (Srinath et al., 2021a,b). It supported user-oriented projects such as the development of privacy question answering assistants (Harkous et al., 2018; Ravichander et al., 2024) and a publicly available web browser extension³ to ease opting out of data collection (Bannihatti Kumar et al., 2020). Further showing its impact beyond research, the corpus enabled or inspired the development of automated privacy compliance tools (Zimmeck et al., 2019; Jain et al., 2023), some now commercially available (e.g., AppCensus⁴ and Google Checks⁵). Additionally, the corpus and subsequent research developments motivated AAAI Symposia in 2016 (Alves-Oliveira et al., 2017) and 2019 (Baldini et al., 2019) as well

¹<https://usableprivacy.org/>

²Google Scholar has multiple redundant entries for the original OPP-115 paper, and merging the three entries known to the authors creates this total.

³<https://addons.mozilla.org/en-US/firefox/addon/opt-out-easy/>

⁴<https://appcensus.io/>

⁵<https://checks.google.com/>

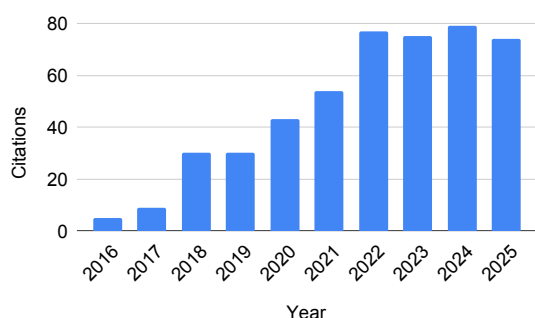


Figure 1: Citations per year to Wilson et al. (2016b), the paper marking the release of the OPP-115 Corpus.

as a 2025 Dagstuhl seminar focused on automating privacy document analysis (Schaub et al., 2025).

This paper reflects on the impact of the OPP-115 Corpus a decade after its release. We describe design decisions that were likely to have contributed to its widespread use, and we summarize research achievements that it seeded. We also share some lessons learned by the research team, hoping to inform future efforts. We conclude with open problems that the corpus surfaced or motivated.

2 Design Decisions

We highlight decisions that encouraged use of the OPP-115 Corpus, which serve as recommendations to future projects to follow similar practices. We acknowledge these choices are not universally practical; larger or more complex corpora, for example, may require radically different approaches.

Format. First, *the corpus was designed to require a minimum amount of technical knowledge to explore and use*. Privacy policy researchers come from a variety of scholarly backgrounds, including law and public policy, where computer programming is not a core skill. Additionally, students are foundational to research labor, and their programming skills are sometimes nascent. Accordingly, the corpus is encoded chiefly in structured text formats, i.e., HTML, XML, and CSV files. All these formats are human readable with some effort, and a web browser and a spreadsheet application are sufficient for guided exploration. Alternatives such as a SQL database would have introduced barriers that discourage marginally skilled researchers. Other features that support a broad user base include the provision of the policy text both in original HTML and in “sanitized” form (i.e., minus HTML tags), illustrating this intermediate step toward annotation,

and the comprehensive explanation of the CSV file columns in the corpus manual.

Enabling reuse. Second, *the practical and legal barriers to accessing the corpus are low*. No payment (or even an email address) is required to download the corpus, contrasting with corpora available via services such as the Linguistic Data Consortium.⁶ Similarly, while version control software is appealing for an audience steeped in software development, it creates a technical barrier; in contrast, the OPP-115 Corpus is downloadable directly via HTTP from the project website⁷. The usage license, inspired by a Creative Commons Attribution-NonCommercial License⁸, welcomes scholarly activities. Additionally, a data exploration website provided visualizations that showed potential users the breadth and depth of the annotations (Cherivirala et al., 2016). The exploration site is no longer available, but Figure 2 shows a screenshot from its interface, where the privacy policy of a website is color-coded by categories of data practices.

Reflecting policy content. Third, *the corpus’s annotation scheme was designed to closely reflect how data practices are described in privacy policies*. Prior projects often focused on answering specific questions about privacy policies or their data practices (Zimmeck and Bellovin, 2014; Reidenberg et al., 2015). These efforts faced challenges due to the ambiguity inherent in privacy policies, with different parts of policy documents potentially contradicting each other or contextual information substantially modifying data practices (Wilson et al., 2016c; Reidenberg et al., 2016). The OPP-115 Corpus’ annotation scheme instead combined a top-down and bottom-up approach: ten categories of data practices were identified based on legal requirements for privacy policies. For each type of data practice, a category-specific attribute set was developed based on input from domain experts (privacy, public policy, and law) and an analysis of how respective data practices were practically described in privacy policy documents in the wild. In the corpus, data practice attributes are further associated with specific text spans (rather than paragraphs) to ground annotations in specific language.

Preserving ambiguity. Fourth, *the corpus is structured to acknowledge the ambiguity and diffi-*

⁶<https://www ldc.upenn.edu/>

⁷<https://usableprivacy.org/data>

⁸<https://creativecommons.org/>

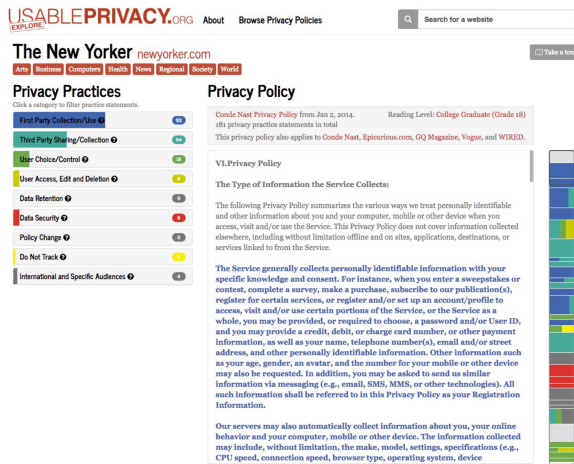


Figure 2: Screen shot from the OPP-115 Corpus data exploration website.

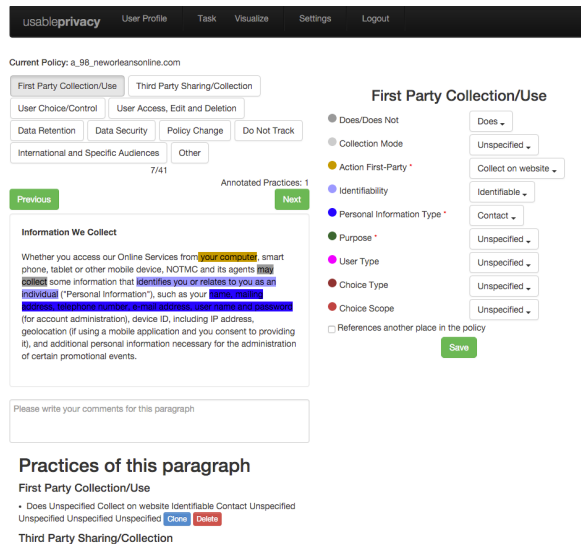


Figure 3: Screen shot of the purpose-built OPP-115 annotation tool.

culty of its data annotation tasks. OPP-115 annotation frames attach to the text via text spans, which render the concept of *inter-annotator agreement* complex and potentially indeterminate. Deciding whether two annotators identified the same underlying data practice, for example, requires several interrelated judgments on text span overlaps and categorical selections. The corpus and documentation mostly avoid claims about *correct* annotations or *gold standard* data. In addition, the corpus’s annotation scheme recognizes “Unspecified” as a valid value for each attribute, to enable annotators to express absence of clear information.

3 Lessons Learned

Although the OPP-115 Corpus has received significant use and attention, some decisions in its creation may have been suboptimal. We identify some of these shortcomings and how they could have been avoided.

The corpus was created with a closed-source, purpose-built annotation tool, leading to obstacles and missed opportunities. Figure 3 shows a screen shot of the annotation tool. One notable strength was its web-based architecture, streamlining the participation of law students at multiple universities while the project was managed from Carnegie Mellon University (Wilson et al., 2016b). Paragraphs from a privacy policy were displayed one at a time in the box at center-left. To create a data practice, an annotator first selected a category (e.g., *First-Party Collection/Use*) and then chose values for attributes (at right) and text spans to justify choices. The *Save* button finalized the data practice, adding it to the list at bottom. The annotator performed this procedure to label all data practices they found in the paragraph, and then used the *Next* or *Previous* buttons to move to another one. It is possible that no one text annotation tool available at the time satisfied all these needs, but annotation tools existed for frame semantics and named entity recognition (Baker, 2008; Marrero et al., 2013). Alternatively, the NLP research community may have welcomed an annotation tool with these features.

Relatedly, *the large size and complexity of the annotation scheme may have limited annotators’ accuracy and inter-annotator agreement.* Creating a data practice required selecting one of ten categories and then selecting values for up to nine category-specific attributes. Some attributes had many possible values; for example, *Personal Information Type* had sixteen and *Action First-Party* had eleven. *Other* and *Unspecified* appear frequently as possible values, giving the annotator flexibility but also reducing their motivation to make complex judgments. Although this annotation scheme appears to produce highly articulate results, the value of the labor is diminished by the likelihood of annotator fatigue and poor inter-annotator agreement (Bayerl and Paul, 2011). The original paper shows the annotators’ frequent lack of consensus through an effort to consolidate their work (§4.2 in Wilson et al. (2016b)); accordingly, the corpus does not commit to a singular *gold standard* set of annotations. An alternative approach could have in-

volved decomposing the annotation procedure into subtasks, such as labeling text spans as privacy-relevant entities and then (as a separate, subsequent task) linking them together to create data practices.

4 Open Problems

We conclude with some research problems that the OPP-115 Corpus revealed and remain fertile for future work. We borrow inspiration from two prior collections of research problems related to privacy policies (Wilson et al., 2016a; Schaub et al., 2025).

Principled consolidation of multiple annotators’ data practices remains an unsolved task. The original paper describes a procedure to identify and combine potentially duplicate data practices created by different annotators (§4.2 in Wilson et al. (2016b)), but this procedure relies chiefly on overlaps between text spans and majority voting for attributes’ values. Those choices render it agnostic to the subject matter of data practices, which can provide a stronger basis for reconciliation. Although consolidation comes with limitations—observations in the original paper suggest that annotators sometimes identify radically different data practices in a paragraph—the expectations of prevalent machine learning paradigms motivate studying this problem.

Reasoning and inference based upon OPP-115 annotations has received little or no attention. The annotation frames are sufficiently expressive to support this activity; for example, *First Party Collection/Use* and *Third Party Sharing/Collection* data practices contain a *Does/Does Not* attribute that provides negation. Data practices in those categories and others contain enough attributes to encode meaningful claims independently of the text spans, easing analysis. Those claims from a given privacy policy can be examined for contradictions, consolidated beyond the boundaries of paragraphs (eliminating a barrier imposed by the OPP-115 annotation tool), and used to create a precise, symbolically-grounded summary of the policy. While large language models (LLMs) are nominally capable of these activities (Rodriguez et al., 2024), symbolic reasoning remains appealing in computational law, where precise provenance of claims is valuable.

Constructing a new corpus with OPP-115-style annotations would provide significant value to the research community. The privacy policies in the OPP-115 Corpus predate many privacy laws (notably GDPR and CCPA, but also others) (Gupta

et al., 2024) that introduced new disclosure and data subject rights requirements (e.g., legal basis of processing; do-not-sell opt-outs, etc.), as well as recent developments in generative artificial intelligence, wearable computing, and immersive reality. Consequently, language models built on the OPP-115 Corpus are likely to perform suboptimally on modern privacy policies. Additionally, a key challenge this brings to light is keeping annotated policy corpora updated over time. Furthermore, most corpora and analyses have focused on English-language privacy policies. While some efforts have analyzed and compared privacy policies across languages, most corpora are either small in terms of numbers of policies annotated (Arora et al., 2022) or the set of analyzed practices is small (Degeling et al., 2019). One exception is a corpus by Moran et al. (2026), although it was LLM-annotated and used machine translations of the OPP-115 Corpus as training data.

Finally, *creating a modernized, general-purpose privacy policy annotation scheme* would benefit privacy policy research. Other works’ privacy policy annotation schemes typically focus on subtopics such as opt-out statements (Bannihatti Kumar et al., 2020) or compliance with specific laws (Arora et al., 2022; Xian et al., 2025). While much of the OPP-115 Corpus annotation scheme generalizes to the present day (e.g., for GDPR (Poplavska et al., 2020)), it falters where it explicitly mentions specific privacy-enhancing technologies or legal provisions. Two examples are *Do Not Track*, which was largely abandoned soon after the corpus was created (Purdy, 2024) and was replaced by “frictionless” opt-out signals as required by CCPA, and *Safe Harbor* provisions, invalidated by the Court of Justice of the European Union in 2015 (Weiss and Archick, 2016) and replaced by the EU-US Data Privacy Framework. A modernized annotation scheme could update items like those or avoid them altogether, focusing instead on slower-changing fundamentals. Retaining the ambitious goal of representing all privacy-related information in a policy would support a wide variety of use cases, across both the analysis of privacy practices and the creation of technologies to support consumer privacy.

5 Conclusion

OPP-115 remains unique, as a corpus of privacy policies with a generalized annotation scheme ap-

plied by skilled annotators. Over the past decade it directly enabled or inspired a variety of projects to make the contents of privacy policies more available to consumers, regulators, privacy scholars, and other parties invested in the privacy landscape. We reflected on the design decisions that contributed to the corpus’ success and lessons learned from its development and use. Finally, we described some open research problems that merit attention by the NLP and privacy research communities.

Acknowledgments

This manuscript is based upon work supported by the National Science Foundation under Grant #2237574. The original research to develop the OPP-115 Corpus was part of a National Science Foundation Frontier project funded by Grant #1330596. This work benefited from Dagstuhl Seminar 25021 “Grand Challenges for Research on Privacy Documents”. The authors wish to acknowledge the law students at Fordham University and the University of Pittsburgh who worked as annotators to make the OPP-15 Corpus possible. Finally, the authors thank all members of the Usable Privacy Policy Project for their contributions.

Limitations

This paper shares reflections on prior research activities, and its stance is comparable to a position paper. It is not intended to present new research results or to be comprehensive on its topics, but it uses supporting evidence from cited literature and the contents of the OPP-115 Corpus.

Ethical Considerations

This paper is reflective in nature, and its ethical implications are limited to the impact of its stance and justifications, which chiefly surround a corpus of annotated privacy policies. No new code or new datasets accompany the paper. Because of that scope, we see minimal or no potential misuse of its contents. The most likely beneficiaries of this paper are other researchers, who may apply the reflections in it to their own work.

References

Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. [PolicyQA: A reading comprehension dataset for privacy policies](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*,

pages 743–749, Online. Association for Computational Linguistics.

Patrícia Alves-Oliveira, Richard G. Freedman, Dan Grollman, Laura Herlant, Laura Humphrey, Fei Liu, Ross Mead, Frank Stein, Tom Williams, and Shomir Wilson. 2017. [Reports on the 2016 AAAI fall symposium series](#). *AI Magazine*, 38(2):86–90.

Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2021. [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.

Siddhant Arora, Henry Hosseini, Christine Utz, Vinayshekhar Bannihatti Kumar, Tristan Dhellemmes, Abhilasha Ravichander, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, Thomas Norton, Thomas Hupperich, Shomir Wilson, and Norman Sadeh. 2022. [A tale of two regulatory regimes: Creation and analysis of a bilingual privacy policy corpus](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5460–5472, Marseille, France. European Language Resources Association.

Collin Baker. 2008. FrameNet, present and future. In *The First International Conference on Global Interoperability for Language Resources*, pages 12–17.

Ioana Baldini, Clark Barrett, Antonio Chella, Carlos Cinelli, David Gamez, Leilani H. Gilpin, Knut Hinkelmann, Dylan Holmes, Takashi Kido, Murat Kocaoglu, William F. Lawless, Alessio Lomuscio, Jamie C. Macbeth, Andreas Martin, Ranjeev Mittu, Evan Patterson, Donald Sofge, Prasad Tadepalli, Keiki Takadama, and Shomir Wilson. 2019. [Reports of the AAAI 2019 spring symposium series](#). *AI Magazine*, 40(3):59–66.

Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, and Norman Sadeh. 2020. Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text. In *Proceedings of The Web Conference 2020*, pages 1943–1954.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Sushain K Cherivirala, Florian Schaub, Mads Schaarup Andersen, Shomir Wilson, Norman Sadeh, and Joel R Reidenberg. 2016. Visualization and interactive exploration of data practices in privacy policies. In *Symposium on Usable Privacy and Security*, pages 3–10.

Lorrie Faith Cranor. 2012. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.*, 10:273.

- Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We value your privacy... Now take some cookies: Measuring the GDPR’s impact on web privacy. In *Network and Distributed Systems Security (NDSS) Symposium*.
- Sonu Gupta, Geetika Gopi, Harish Balaji, Ellen Poplavska, Nora O’Toole, Siddhant Arora, Thomas Norton, Norman Sadeh, and Shomir Wilson. 2024. [Creation and analysis of an international corpus of privacy laws](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4092–4105, Torino, Italia. ELRA and ICCL.
- Hamza Harkous, Kassem Fawaz, Rémi Lebrete, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 531–548.
- Akshath Jain, David Rodriguez, Jose M Del Alamo, and Norman Sadeh. 2023. Atlas: Automatically detecting discrepancies between privacy policies and privacy labels. In *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 94–107. IEEE.
- Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. 2013. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.
- Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *ISJLP*, 4:543.
- Abraham Mhaidli, Selin Fidan, An Doan, Gina Herakovic, Mukund Srinath, Lee Matheson, Shomir Wilson, and Florian Schaub. 2023. Researchers’ experiences in analyzing privacy policies: Challenges and opportunities. *Proceedings on Privacy Enhancing Technologies*.
- Marcos Moran, David Rodriguez, Luka Nenadic, Norman Sadeh, and Jose M. Del Alamo. 2026. [Enabling multilingual privacy policy audits: Large-scale analysis of spanish mobile apps](#). *arXiv preprint*.
- Ellen Poplavska, Thomas B Norton, Shomir Wilson, and Norman Sadeh. 2020. From prescription to description: Mapping the gdpr to a privacy policy corpus annotation scheme. In *Legal Knowledge and Information Systems: JURIX 2020: The Thirty-Third Annual Conference, Brno, Czech Republic, December 9–11, 2020*, pages 243–246. SAGE Publications Pvt. Ltd 1 Oliver’s Yard, 55 City Road, London, EC1Y 1SP.
- Kevin Purdy. 2024. [Back where it started: “Do Not Track” removed from Firefox after 13 years](#). *Ars Technica*. Accessed: March 12, 2026.
- A. Ravichander, I. Yang, R. Chen, S. Wilson, T. Norton, and N. Sadeh. 2024. Incorporating taxonomic reasoning and regulatory knowledge into automated privacy question answering. In *Proceedings of the International Conference on Web Information Systems Engineering (WISE)*, pages 444–460.
- Abhilasha Ravichander, Alan W Black, Thomas Norton, Shomir Wilson, and Norman Sadeh. 2021. [Breaking down walls of text: How can NLP benefit consumer privacy?](#) In *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 4125–4140, Online. Association for Computational Linguistics.
- Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. [Question answering for privacy policies: Combining computational and legal perspectives](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958, Hong Kong, China. Association for Computational Linguistics.
- J. R. Reidenberg, T. D. Breaux, L. F. Cranor, B. French, A. Grannis, J. T. Graves, F. Liu, A. M. McDonald, T. B. Norton, N. C. Russell, N. Sadeh, and F. Schaub. 2015. Disagreeable privacy policies: Mismatches between meaning and users’ understanding. *Berkeley Technology Law Journal*, 30:39–88.
- Joel R Reidenberg, Jaspreet Bhatia, Travis D Breaux, and Thomas B Norton. 2016. Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies*, 45(S2):S163–S190.
- David Rodriguez, Ian Yang, Jose M. Del Alamo, and Norman Sadeh. 2024. [Large language models: a new approach for privacy policy analysis at scale](#). *Computing*, 106(12):3879–3903.
- Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Aleecia M McDonald, Joel R Reidenberg, Noah A Smith, Fei Liu, N Cameron Russell, Florian Schaub, and Shomir Wilson. 2013. The usable privacy policy project. In *Technical Report CMU-ISR-13-119*. Carnegie Mellon University.
- Florian Schaub, Christine Utz, Shomir Wilson, and Lu Xian. 2025. Grand challenges for research on privacy documents (Dagstuhl Seminar 25021). *Dagstuhl Reports*, 15(1):1–32.
- Mukund Srinath, Soundarya Nurani Sundareswara, C. Lee Giles, and Shomir Wilson. 2021a. [Privaseer: A privacy policy search engine](#). In *Web Engineering: 21st International Conference, ICWE 2021, Biarritz, France, May 18–21, 2021, Proceedings*, page 286–301, Berlin, Heidelberg. Springer-Verlag.
- Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021b. [Privacy at scale: Introducing the PrivaSeer corpus of web privacy policies](#). In *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. Association for Computational Linguistics.

Martin A. Weiss and Kristin Archick. 2016. [U.S.-EU Data Privacy: From Safe Harbor to Privacy Shield](#). Technical Report R44257, Congressional Research Service. CRS Report.

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 with language technologies: Progress and challenges. In *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)*.

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. 2016b. [The creation and analysis of a website privacy policy corpus](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, Berlin, Germany. Association for Computational Linguistics.

Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A. Smith, and Frederick Liu. 2016c. [Crowdsourcing annotations for websites' privacy policies: Can it really work?](#) In *Proceedings of the 25th International Conference on World Wide Web, WWW '16*, page 133–143, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Lu Xian, Van Hong Tran, Lauren Lee, Meera Kumar, Yichen Zhang, and Florian Schaub. 2025. [Layered, overlapping, and inconsistent: A large-scale analysis of the multiple privacy policies and controls of u.s. banks](#). In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security, CCS '25*, page 3177–3191, New York, NY, USA. Association for Computing Machinery.

Sebastian Zimmeck and Steven M Bellovin. 2014. Privee: An architecture for automatically analyzing web privacy policies. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 1–16.

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