Multilingual Scraper of Privacy Policies and Terms of Service

Karel Kubicek
karel.kubicek@inf.ethz.ch

David Bernhard
dbernhard@ethz.ch

Stefan Bechtold
sbechtold@ethz.ch

1 INTRODUCTION

Privacy policies (*policies)* document how firms collect, store, and use users' personal data. They are of significant interest in many studies examining data collection practices of websites, mobile apps, and other services [1, 5, 6, 8, 11, 12, 14]. Terms of service (or terms and conditions, *terms*) are legal contracts between the consumer and the service, and therefore are also a focus of legal studies [2, 3].

Empirical studies of policies and terms rely on a corpus of such documents. While some datasets exist, they either provide documents from a single timestamp [14] or focus on historical data [1]. We are not aware of any continuously operating project collecting policies and terms. Consequently, many researchers develop their own scraping tools, usually with the following limitations.

(1) Restriction to English-speaking websites, leading to the systematic understudy of non-English websites, as discussed by Mhaidli et al. [11, Sec. 4.4.2].
(2) Focus on single measurements, ignoring the evolution over time, which is essential for assessing trends caused by privacy regulations and technologies.
(3) Non-representative website sampling methods that do not reflect real users' browsing habits, as shown by Ruth et al. [13]. Additionally, the samples often overrepresent the US population compared to other countries.

Several prior works have attempted to address some of these limitations. Hosseini et al. [6] presented a unified privacy policy scraper, but it is limited to English and German and does not perform periodic crawls. Amos et al.[1] inspected the history of policies using Web Archive, but their sample is limited to archived websites and the project concluded in 2021. Degeling et al. [4], Linden et al. [10], and Hosseini et al. [7] studied the evolution of policies during GDPR adoption, but these works have also been discontinued.

We developed and deployed a scraper addressing these issues. It supports 37 languages (see Appendix A), enabling future studies of underrepresented countries. Our deployment focuses on long-term data collection, crawling nearly 1 million websites monthly for five years. We use the Chrome UX Report (*CrUX*) list for sampling, which, according to Ruth et al. [13], closely represents actual user browsing behavior. Our sample is diverse in selection of countries and popularity levels. In addition to policies, we also collect terms, and other legal documents.

2 ARCHITECTURE

Our crawler utilizes a real browser through the `Selenium` library. While this choice increases computation time, it significantly reduces the chances of detection as a bot,[1] minimizing bias in the scraped policies and terms. The crawler navigates websites using annotated keywords for various page types, matching text with all page links. If the target document is not found on the index page, the crawler navigates to login and registration pages, which refer to policies and terms more often, or randomly browses within the starting domain. If navigation fails to locate the desired documents, we use search engines (`startpage.com` or `duckduckgo.com`), restricting searches to the target domain.

Upon reaching a policy or terms page, we extract the text body using the `readability` library. We classify the document using a machine-learning model (see Appendix A), storing clear text in a database and raw HTML on disk if it is the desired document.

[1]We employ multiple methods to reduce bot detection. See [9, Sec. A.1] for the full list.

3 CRAWLING LIST

To capture the evolution of policies and terms and trends in specific website populations, we created static and dynamic samples, both sampled using similar strategies. The static list, sampled once from CrUX 2023-12, contains 502 612 websites to be crawled for five years. The dynamic list is sampled monthly from the latest CrUX release. We crawl the union of these lists, currently about 800k websites, which is expected to increase as the 2023-12 CrUX list gradually outdates.

CrUX groups websites based on popularity in specific countries, with popularity buckets of 1k, 5k, 10k, 50k, 100k, 500k, 1M, and 5M. The list of countries of interest is in Appendix A. To obtain a representative sample, we randomly sample 5k websites (or bucket size for 1k and 5k buckets) from each bucket. We take the union of all samples, reducing the expected 870k websites to roughly 500k due to overlaps.

4 LONG-TERM SUPPORT

Our regular crawls have begun in January 2024. To operate until the end of 2028, the system is tuned for minimal maintenance, using continuous integration and development for autonomous updates. Based on almost a year of testing, we developed monitoring that reports errors to the responsible team. If no problems are observed, it sends an overview email with monthly statistics. We expect to collect over 2 TB of extracted policies and terms texts stored in a PostgreSQL database and 5 TB of compressed HTML dumps. The operation is supported by Professor Stefan Bechtold at ETH Zurich.

5 ACCESS

We are open to providing access to interested parties for the following scopes.

(1) GitLab interface for browsing the policies and terms and observing changes over time.
(2) Database access for performing large-scale studies.
(3) Access to individual HTML documents via an API.

Please indicate your interest on the project page at the following link https://karelkubicek.github.io/post/pptc (the QR code next to the title leads to it as well). You can also sign up for a newsletter with monthly updates or join the mailing list for the research community around empirical analysis of policies and terms.

Table 1: Correctness of policies and terms detection depending on keyword-based crawler navigation or search engine.

Documents	Detection method	Correct doc.	Wrong doc.
Policies	Keyword matching	84.2% (101)	15.8% (19)
	Search engine	86.7% (104)	13.3% (16)
Terms	Keyword matching	84.2% (101)	15.8% (19)
	Search engine	66.7% (80)	33.3% (40)

Table 2: Evaluation of found policies and terms.

Observation	Policies	Terms
No document found, none present	24.0% (66)	33.1% (91)
No document found, document present	21.1% (58)	25.1% (69)
Found document, none present	0.7% (2)	1.1% (3)
Found wrong document	5.8% (16)	4.7% (13)
Found correct document	48.4% (133)	36.0% (99)

6 RESULTS

Our crawler successfully loads 97.8% of websites. We found a policy on 48.4% of these websites, namely 389 358 were found using keyword-based navigation and 10 473 using search. We also found terms on 34.7% of successfully loaded websites, namely 281 366 using navigation and 5313 using search. Table 1 presents the correctness of the found documents depending on the detection method. Also, Table 2 provides an end-to-end evaluation of detection rate and presence of the policies and terms.

REFERENCES

[1] 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* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 2165–2176. https://doi.org/10.1145/3442381.3450048

[2] Yannis Bakos, Florencia Marotta-Wurgler, and David R Trossen. 2014. Does anyone read the fine print? Consumer attention to standard-form contracts. *The Journal of Legal Studies* 43, 1 (2014), 1–35.

[3] Giuseppe Dari-Mattiacci and Florencia Marotta-Wurgler. 2022. Learning in Standard-Form Contracts: Theory and Evidence. *Journal of Legal Analysis* 14, 1 (2022), 244–314. https://doi.org/10.1093/jla/laad001

[4] 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 *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019)*. The Internet Society, San Diego, CA, USA, 1–15. https://doi.org/10.14722/ndss.2019.23378

[5] Hamza Harkous, Kassem Fawaz, Rémi Lebret, 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)*. USENIX Association, Baltimore, MD, 531–548. https://www.usenix.org/conference/usenixsecurity18/presentation/harkous

[6] Henry Hosseini, Martin Degeling, Christine Utz, and Thomas Hupperich. 2021. Unifying privacy policy detection. *Proceedings on Privacy Enhancing Technologies* 2021 (2021), 480–499. Issue 4. https://doi.org/10.2478/popets-2021-0081

[7] Henry Hosseini, Christine Utz, Martin Degeling, and Thomas Hupperich. 2024. A Bilingual Longitudinal Analysis of Privacy Policies Measuring the Impacts of the GDPR and the CCPA/CPRA. *Proceedings on Privacy Enhancing Technologies* 2024 (2024), 434–463. Issue 2. https://doi.org/10.56553/popets-2024-0058

[8] Michael Kretschmer, Jan Pennekamp, and Klaus Wehrle. 2021. Cookie Banners and Privacy Policies: Measuring the Impact of the GDPR on the Web. *ACM Trans. Web* 15, 4, Article 20 (jul 2021), 42 pages. https://doi.org/10.1145/3466722

[9] Karel Kubicek, Jakob Merane, Ahmed Bouhoula, and David Basin. 2024. Automating Website Registration for Studying GDPR Compliance. In *Proceedings of the ACM Web Conference 2024*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3589334.3645709

[10] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2020. The privacy policy landscape after the GDPR. *Proceedings on Privacy Enhancing Technologies* 2020, 1 (2020), 47–64. https://doi.org/10.2478/popets-2020-0004

[11] 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* 2023 (2023), 287–305. Issue 4. https://doi.org/10.56553/popets-2023-0111

[12] Christian Peukert, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer. 2022. Regulatory Spillovers and Data Governance: Evidence from the GDPR. *Marketing Science* 41, 4 (2022), 746–768. https://doi.org/10.1287/mksc.2021.1339

[13] Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric. 2022. Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists. In *Proceedings of the 22nd ACM Internet Measurement Conference* (Nice, France) *(IMC '22)*. Association for Computing Machinery, New York, NY, USA, 374–387. https://doi.org/10.1145/3517745.3561444

[14] 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, et al. 2016. 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)*. Association for Computational Linguistics, Berlin, Germany, 1330–1340. https://doi.org/10.18653/v1/P16-1126

A APPENDIX

This summary our system's capabilities is not exhaustive. See Kubicek et al. [9, Section 3 and Appendix A] for details.

Supported languages. Our crawler supports 37 languages, with most keywords translated by native or proficient speakers who observed multiple websites prior to the translation. These languages are: Bulgarian, Bosnian, Catalan, Czech, Welsh, Danish, German, Greek, English, Spanish, Estonian, Basque, Finnish, French, Galician, Croatian, Hungarian, Icelandic, Italian, *Luxembourgish*, Lithuanian, Latvian, Macedonian, *Maltese*, Dutch, Norwegian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Albanian, Serbian, Swedish, Turkish, and Ukrainian. Two of these languages are not supported by multilingual BERT, which we use for document classification, so languages in italics have limited support.

Selected countries and maximal ranks. We sample CrUX for the following list of countries. Each country has a maximal rank in which there are websites available, we denote this in parentheses. United States (5M), Great Britain (5M), Switzerland (500k), Iceland (50k), Norway (500k), Lichtenstein (50k), Turkey (1M), Russia (1M), France (1M), Germany (1M), Austria (500k), Belgium (500k), Bulgaria (500k), Croatia (100k), Cyprus (50k), Czechia (500k), Denmark (500k), Estonia (100k), Finland (500k), Greece (500k), Hungary (500k), Ireland (500k), Italy (1M), Latvia (100k), Lithuania (100k), Luxembourg (100k), Malta (50k), Netherlands (1M), Poland (1M), Portugal (500k), Romania (500k), Slovakia (500k), Slovenia (500k), Spain (1M), and Sweden (500k).

Policies and terms classification. To classify policies and terms, we train two binary multilingual distilled BERT models: one on 415 positive and 133 negative samples of policies, and another on 273 positive and 810 negative samples of terms. These multilingual datasets were labeled based on detected documents by our crawler, representing the actual distribution observed. The models achieve 93.2% and 92.3% accuracy for policies and terms, respectively. In comparison, a model using structure from [1] achieved 80.0% accuracy and it is limited to policies only.