Multilingual Scraper of Privacy Policies and Terms of Service

David Bernhard
dbernhard@ethz.ch
ETH Zurich

Luka Nenadic
luka.nenadic@gess.ethz.ch
ETH Zurich

Stefan Bechtold
sbechtold@ethz.ch
ETH Zurich

Karel Kubicek
karel.kubicek@inf.ethz.ch
ETH Zurich, Inria

ABSTRACT

Websites' privacy policies and terms of service constitute valuable resources for scholars in various disciplines. Nonetheless, there exists no large, multilingual database collecting these documents over the long term. Therefore, researchers spend a lot of valuable time collecting them for individual projects, and these heterogeneous methods impede the reproducibility and comparability of research findings. As a solution, we introduce a long-term scraper of privacy policies and terms supporting 37 languages. We run our scraper on a monthly basis on 800 000 websites, and we publish the dataset for the twelve crawls in 2024. Our manual evaluation of the end-to-end extraction of the documents demonstrates F1 scores of 79% for privacy policies and 75% for terms of service in five sample languages (English, German, French, Italian, and Croatian). We present several broad potential applications of our database for future research.

CCS CONCEPTS

• **Applied computing** → **Law**; • **Security and privacy** → **Privacy protections**; • **Information systems** → *World Wide Web*.

KEYWORDS

Web scraping, dataset, privacy, privacy policies, terms of service

ACM Reference Format:
David Bernhard, Luka Nenadic, Stefan Bechtold, and Karel Kubicek. 2025. Multilingual Scraper of Privacy Policies and Terms of Service. In *Symposium on Computer Science and Law (CSLAW '25), March 25–27, 2025, München, Germany*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3709025.3712215

1 INTRODUCTION

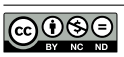
The Internet as we know it today has enabled countless mass markets that connect millions of firms with billions of customers around the globe. A traditional legal tool to structure the relationship between a firm and its customer is a contract. In mass markets, firms typically enter into contractual relationships with all their customers. While mass-market standard contracts were already important before the emergence of the World Wide Web[1], online commerce has made mass-market contracts a truly pervasive phenomenon [39]. In online commerce, firms faced many novel legal challenges such as online purchase and delivery, permissible website use, or the collection and processing of consumer data. At the same time, firms needed to comply with existing and emerging legislation in domains such as privacy, copyright, or consumer protection law.

Today, we expect firms to have addressed these challenges in extensive legal documents on their websites. Privacy policies (*policies)* document how firms collect, store, and use our personal data. Terms of service (*terms*) cover a wide array of legal topics, ranging from permissible website use, payment terms, and intellectual property rights to limitations of liability, governing law, and jurisdiction.

Policies and terms provide a wealth of information for researchers. Although consumers are not paying much attention to online standard form contracts [4, 38], researchers can fill the gap by uncovering interesting provisions in policies and terms. For example, a typical research question asks whether a particular piece of legislation has impacted both the policies that are subject to it and those that are not (in the context of the European Union's General Data Protection Regulation (*GDPR*), see [10, 16]).

Accessing policies and terms can be a challenging endeavor, in particular if researchers need to access them at scale. This often leads researchers to settle for small, hand-crafted datasets. This is particularly true for legal scholars who often lack the technical expertise or institutional support to develop an automated web crawling infrastructure. Focusing on small datasets has its costs. On the one hand, small sample sizes can lead to biased results. Focusing, for example, on only top-ranked websites ignores the diversity and the long tail of the web[2], thereby potentially missing different trends for less popular websites. On the other hand, even if one succeeds at building a scraping infrastructure for individual projects, this process is not efficient: valuable time and resources are not allocated to actual research and the replicability and comparability of any research findings are impeded.

Therefore, a readily available dataset of policies and terms would be highly beneficial to the research community. This has also been reported by Mhaidli et al. [33], who interviewed members of the research community studying privacy policies. They identify several challenges, based on which we formulate the following criteria for dataset collection:



CSLAW '25, March 25–27, 2025, München, Germany

ACM ISBN 979-8-4007-1421-4/25/03.
https://doi.org/10.1145/3709025.3712215

[1] Back in 1971, Slawson [44] estimated that 99% of commercial contracts produced in advanced economies are standard form contracts.
[2] It is estimated that there are more than 1 billion websites in the world, with 18% of them being active [37].

(1) *Scale*: The dataset should contain terms and policies of a very large number of websites to address the issues set forth in the preceding paragraph and by Mhaidli et al. [33, Sec. 4.1.3]. While a manual analysis of such text corpora might have been practically impossible in the past, recent advances in natural language processing and machine learning enable automating large parts of the process.
(2) *Multilingualism*: The dataset should not be limited to the English language. As reported by Mhaidli et al.[33, Sec. 4.4.2], the current research on terms and policies is English- and Western-centric. Such limitation could obfuscate heterogeneity across different languages and regions. Furthermore, certain empirical legal studies require data in a specific language, when, for example, a researcher wants to examine the effects of a policy shift in consumer protection law on terms and policies in Italy.
(3) *Long-Term Support*: To enable longitudinal analyses, the dataset should be dynamic and entail different versions of policies and terms over time (Mhadli et al.[33, Sec. 5.1.4]). Data gathering should be based on an infrastructure with streamlined maintenance and minimal support effort. This is essential for assessing trends caused by emerging regulations and technologies.
(4) *Both terms and policies*: The dataset should contain both policies and terms of websites. By supporting both policies and terms, one could, for example, study how they interact.
(5) *Standardized format*: To facilitate analysis, the dataset should consist of policies and terms in a standardized format.

We are not aware of any continuously operating project that aspires to these principles. As a solution, we introduce a scraper that visits around 800 000 websites per month, scraping them for policies and terms in 37 languages and storing their content in a database.

The rest of the paper is structured as follows. Section 2 presents related work. Section 3 describes the scraper's architecture in detail. Section 4 presents our results, while Section 6 envisions potential future work using our database. Finally, Section 7 concludes.

2 RELATED WORK

2.1 Legal Research Using Policies and Terms

Both policies and terms have been at the heart of a plethora of studies. As for policies, Mhaidli et al. [33] have synthesized many use cases for researchers across various disciplines. Policies are used to examine the data collection practices of websites, mobile apps, and other services [2, 19, 20, 31, 48]. These practices are often evaluated against the backdrop of regulatory developments [11, 23, 26, 27, 36, 47, 49] such as the introduction of the European Union's GDPR and its potential spillover effects in other jurisdictions [9, 10, 16, 40], or the California Privacy Protection Act [46]. Other researchers have studied the usability and readability of policies [1, 2, 5, 15, 25, 30].

Albeit at a much lesser extent than policies, terms have also served as a basis for several legal studies. One area of focus, similarly to policies, concerns the readability of terms [6, 43]. Other studies have covered the (lack of) attention that users pay to terms [4, 38]. Scholars have further studied change and innovation of terms in software license agreements [8, 32]. Another area of research pertains to uncovering unfair clauses in terms [24, 28, 29, 34]. Finally, Kolt [22] examined the suitability of a large language model to read and understand terms.

2.2 Existing Policy and Terms Corpora

Empirical studies of policies and terms rely on a corpus of such documents. While various datasets exist, they do not fulfill all the criteria set forth in the Introduction. Regarding policies, the Usable Privacy Policy Project [42] provides several useful datasets. Their most famous dataset, the OPP-115 Corpus [48], provides 115 English policies with manual fine-grained annotations. While this dataset proved to be very useful for training [13] and evaluating [18] large language models in the legal field, it is small in size, lacks multilingual support, and is limited to a single timestamp.

Larger datasets include the Princeton-Leuven Longitudinal Corpus of Privacy Policies [2], the PrivaSeer Corpus of Web Privacy Policies [45], MAPS (for app policies on the Google Play Store) [50], and the dataset of Wagner [47]. All these datasets are limited to the English language. Furthermore, only the Princeton-Leuven Longitudinal Corpus of Privacy Policies and Wagner's datasets enable longitudinal analyses; but they are not suited for analyzing future developments, as both datasets are no longer being updated. Other discontinued databases for English policies include Linden et al. [27] and Hosseini et al. [21]. Both datasets were created with the specific purpose to assess the effects of the GDPR on policies.

Many scholars deplore the ensuing lack of research on non-English policies [3, 7, 11, 33]. For that reason, some bilingual datasets were created for English and German [3, 20] as well as English and Italian [7]. For their research on the effects of the GDPR, Degeling et al. [11] even scrutinized policies in 24 different languages. However, all these datasets only offer an individual timestamp of policies. We would like to note the commendable efforts of Hosseini et al. [20] to unify the policy detection process in order to make different studies in the field more comparable.

When it comes to terms, relevant corpora are scarce. In order to train an algorithm that detects unfair clauses, Lippi et al. [29] collected and annotated 50 English terms of online platforms. To assess the performance of GPT-3 on ten yes or no questions, Kolt [22] collected the terms of the 20 most-visited U.S. websites. Samples et al. [43] created a corpus of 75 terms of smartphone-based social platforms to examine the evolution of their linguistic characteristics over time. Drawzeski et al. [14] presented a corpus of 25 terms annotated in four languages (English, German, Italian, and Polish) each. Finally, while Galassi et al.[17] cover both policies and terms in English and German, they only inspect five documents each. All of these corpora suffer from a low sample size, a lack of long-term support, and —for the most part— a focus on solely English terms.

3 ARCHITECTURE

To address the limitations of currently existing datasets, we developed and deployed a multilingual scraper of policies and terms. It supports 37 languages (see Appendix A), enabling future studies of underrepresented languages and countries. Our deployment focuses on long-term data collection, scraping around 800 000 websites monthly for several years. We sample websites from the Chrome

UX Report list (*CrUX*) that closely represents actual user browsing behavior. Our sample is diverse in selection of countries and popularity levels of websites.

To be able to scrape 800 000 websites per month, we engineered the following efficient architecture depicted in Fig. 1. Since the project utilizes Docker containers for deployment and scaling, it could potentially be executed on multiple devices simultaneously. Our current deployment utilizes a single computer with a general-purpose AMD Ryzen 9 7950X processor. The monthly scraping procedure can be summarized as follows:

(1) First, each scraper container fetches a "job," i.e., a website it should scrape, from the database. (2) Then, it visits the website in a Chrome browser and navigates to pages likely containing the policy and terms. When the policy or terms page is loaded, it extracts the document in Markdown clear-text. (2a) In case a policy or terms was not found, the scraper uses a search engine to attempt finding the desired document. (3) To check that the page contains the desired document, the scraper calls the classifier to estimate the probability of the text being the website's policy or terms. If the document is not a policy or terms, step (2) is iterated. (4) Finally, the scraper stores the results in the database and the raw HTML code of the page on the disk.

To scrape all websites within one month, we deploy 44 scraper containers in parallel, but only a single classifier. We use containers because they start faster and use much less resources than virtual machines while still providing reasonable isolation. In our deployment that fully utilizes 32 threads and 64 GB of memory, the scraper can browse all 800 000 websites within around three weeks, scraping at approximately 26 websites per minute.

3.1 Scraper

Our scraper utilizes a real Chrome browser instrumented via the `Selenium` library. While this significantly increases computation time, using a full browser, namely the Undetected Chromedriver project[3], minimizes bias in the scraped policies and terms in two ways. The scraper represents better how typical users observe the website and also reduces the likelihood of being detected as a bot. In Appendix A, we list further measures we implemented to evade bot detection. Additionally, each scraper uses a VPN with a static IP address to access the websites. To simulate users from an EU country, we are using a VPN with six IP addresses of a German university, with the scraper generating only between 2 and 8 MB/s of traffic. Note that the used IP addresses are not publicly classified as a VPN by `ipinfo.io`, so that our traffic likely remains unblocked.

Each month, before visiting websites, the browser starts with a custom browser profile in order to update the scraper's browser and its extensions. *Inter alia*, we use the uBlock Lite extension to skip interactions with cookie notices and other popups blocked by the EasyList Cookie List list and to reduce traffic by blocking advertising.

After fetching the target URL as a job from the database, the scraper visits the index page. After loading each page, the scraper inspects all its hyperlinks, ranking them according to our heuristic in order of links most likely leading to the policy and terms. The scraper primarily searches for the policy and terms links, then for registration or login links, and, if none of the former are available, for general links. This heuristic detects keywords in the URL or in the visible text associated to the link. The keywords were curated by native- or proficient-speakers for the majority of the supported languages.[4]

If the navigation still fails to locate the policy and terms, the scraper queries search engines (`startpage.com` or `duckduckgo.com`), with a search restricted to the target domain, using the keywords for policies and terms. When these searches still yield no result, we attempt to load the most used URL paths for policy/term pages in the language of the given website.[5]

Upon reaching a potential policy or terms page, we extract the text body using the `Readability` library. We classify all potential documents using a machine-learning model and store them as clear-text in the database (including the metadata presented in Section 3.4), even if they did not meet the threshold to be classified as the desired type. We also store the raw HTML on the disk if it concerns the potential policy or terms.

[3]https://github.com/ultrafunkamsterdam/undetected-chromedriver/

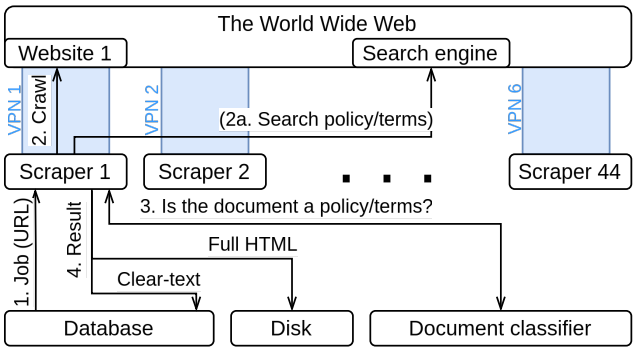


Figure 1: Architecture of the scraping infrastructure.

3.2 Classifier

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.[6] These multilingual datasets were labeled based on detected documents by our scraper, representing the actual distribution observed. The models achieve 93.2% and 92.3% accuracy for policies and terms on the validation dataset, respectively. In comparison, a model using the structure from [2] achieved 80.0% accuracy and is limited to policies only.

3.3 Website Selection

With the capability to scrape around 800 000 websites per month, we cannot scrape the entire internet, but only a sample of interest. We decided to use the CrUX list, because Ruth et al. [41] have observed that it is the most representative of actual user browsing

[4]Taking English policies as an example, the keywords include "privacy [|policy|notice|statement]". This simplified regular expression notation matches the word privacy alone or followed by policy, notice, or statement. Note that longer matches are given a higher score.

[5]For example `/privacy` or `/policies/terms-of-service` for English websites.

[6]The training data was sampled randomly from a monthly crawl.

statistics. CrUX groups websites based on their popularity in specific countries, with popularity buckets of 1k, 5k, 10k, 50k, 100k, 500k, 1M, and 5M.

Since CrUX contains about 23 million websites and is constantly changing, we use a static list, which was sampled once from the CrUX list of December 2023 and contains 502 612 websites. The static list is complemented by a dynamic part, which is resampled monthly from the current CrUX list. We scrape the union between the static list and the dynamic part, currently around 800k websites. This number will gradually increase as shown in Table 1, because the overlap of the static list and the dynamic part shrinks each month.

The countries of interest and their maximal popularity buckets are listed in Appendix A. To obtain a representative sample, we randomly sample 5k websites (or the maximum bucket size for 1k and 5k buckets) from each bucket up to the country-specific maximal bucket. We retain all popularity information for each website in our database.

Table 1: Increasing size of the scraping list.

Month	Number of websites
2024-04	792 614
2024-05	796 574
2024-06	798 650
2024-07	800 237
2024-08	801 949
2024-09	802 010
2024-10	804 693
2024-11	815 136
2024-12	813 172

3.4 Database and Disk Storage

Scraper containers communicate with a central PostgreSQL database to fetch jobs and store results. A simplified entity relationship model representing the database scheme is presented in Fig. 2.

At the beginning of each month, an initialization script samples domains to scrape as described in Section 3.3 and creates jobs for all of them. Each job contains the domain (`crux_url`) and its metadata, such as popularity and timestamps for the start and end of the scrape. Once a job is started, the scraping state is inserted as an entry in the `website_crawls` table. When the scraper identifies a policy page, it stores the classification probability along with the identification source (i.e., by the navigation heuristic, search engine, or popular path) in the `crawl_to_privacy` table. This table also refers to the `urls` and `contents` tables that store the extracted clear-text of the policy. The `contents` table saves storage by preventing the creation of duplicate policies using hashes of existing entries. This would otherwise happen often, as a lot of policies remain unchanged over many months. The same procedure analogously applies to terms.

Each month the database grows by approximately 6 GB, mostly stemming from the clear-text policies and terms. We also store the raw HTML, logs, and clear-text documents on the disk, which in the compressed form occupies around 75 GB per month.

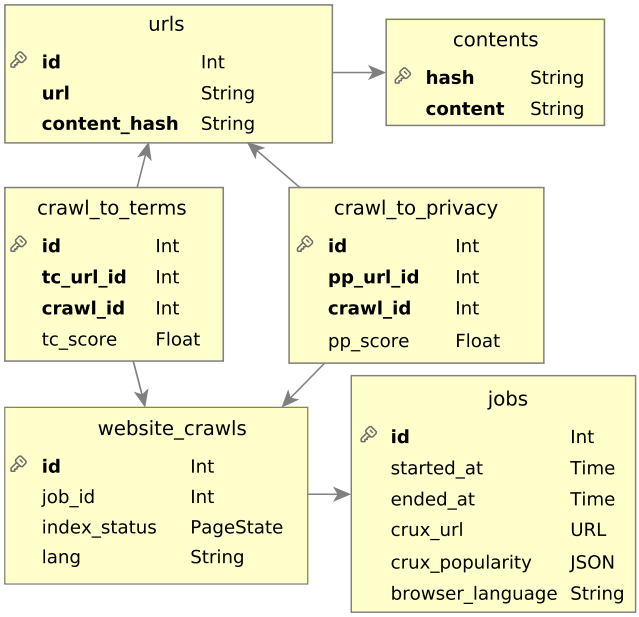


Figure 2: ERM diagram of the simplified database scheme.

3.5 Long-term support

Our regular scraping has begun in October 2023. To operate over the next years, the system is tuned for minimal maintenance, using continuous integration and development for autonomous updates. Based on almost a year of testing, we developed a monitoring system that reports errors to the responsible team. If no problems are encountered, the system sends an overview e-mail with monthly statistics.

4 RESULTS

To assess the performance of our scraper, we have randomly sampled 100 policies and terms each for English, German, French, Italian, and Croatian. The first four languages were selected due to their high likelihood of being relevant for future studies, while Croatian was chosen to gather a first indication of performance for less common languages. To ensure that we would obtain a representative sample, we randomly selected 25 websites from each of the following CrUX popularity buckets: (1) 1k, (2) 5k and 10k, (3) 50k and 100k, as well as (4) 100k and all higher CrUX buckets.

Two annotators assessed the accuracy of the scraper's output. To ensure consistency, we created a codebook (see Appendix A) and the annotators commented edge cases. Each website's most likely result for policy and terms, as determined by our scraper, was annotated into one of the following five cases:

(1) *Wrong: document found, none present*: The scraper found a document, whereas the website actually does not have this type of document. For example, the scraper identified terms, although the website does not have terms.
(2) *Wrong: wrong document found*: Either the scraper found some document, which is not the actual document, or the content extracted by the scraper does not match the actual document, e.g., because the first paragraph is missing.

(3) *Wrong, correct document not found*: The scraper did not find the document, which is actually available on the website. For example, the scraper erroneously reported that the website does not have a policy.
(4) *Correct: no document*: The scraper did not find any document and the website truly does not have this type of document.
(5) *Correct: document found*: The scraper found the correct document and the contents match.

Figs. 3 and 4 show the annotation results for policies and terms, respectively. Furthermore, in Table 2 and Table 3, we report the scraper's metrics for accuracy, recall, precision, and F1 scores related to policies and terms separately. For evaluation purposes, we considered "Correct: document found" to be a true positive, and both "Wrong: document found, none present" as well as "Wrong: wrong document found" to be false positives.

Overall, we can see that the scraper performs well for the four common languages and even delivers satisfactory results for Croatian. We are especially satisfied with the recall rates in English, German, French, and Italian, meaning that our database is likely to contain a website's policy or terms if such are available. The recall rates could even be improved if we decided to consider all potential instances of policies and terms scraped and not only the most likely one. Note that one could also fine-tune performance by either setting different thresholds with the probabilities provided by our classifier or using their own classifier on the scraped data. Nonetheless, the results for Croatian show that less popular languages will likely exhibit worse performance.

Table 2: Policy classification metrics.

	English	German	French	Italian	Croatian	Total
Accuracy	0.73	0.7	0.75	0.72	0.63	0.71
Recall	0.88	0.78	0.81	0.91	0.51	0.79
Precision	0.78	0.86	0.77	0.73	0.71	0.78
F1 score	0.83	0.81	0.79	0.81	0.59	0.79

Table 3: Terms classification metrics.

	English	German	French	Italian	Croatian	Total
Accuracy	0.73	0.76	0.73	0.83	0.69	0.75
Recall	0.81	0.82	0.73	0.8	0.7	0.78
Precision	0.75	0.79	0.71	0.74	0.6	0.73
F1 score	0.78	0.81	0.72	0.77	0.64	0.75

Finally, we evaluated all 155 false positives and present the encountered error types in Table 4. In the majority of the cases, the scraper retrieved the wrong document, such as the cookie notice instead of the policy or the imprint instead of the actual terms. The scraper sometimes experienced technical issues by being blocked or not being able to extract the markdown text or encountered unusual websites, where, e.g., the terms were scattered across several pages.

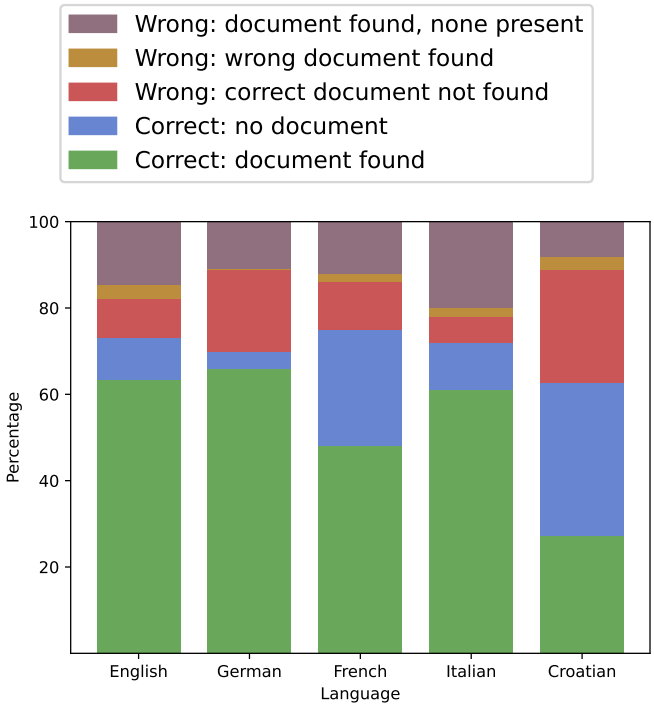


Figure 3: Evaluation of found policies.

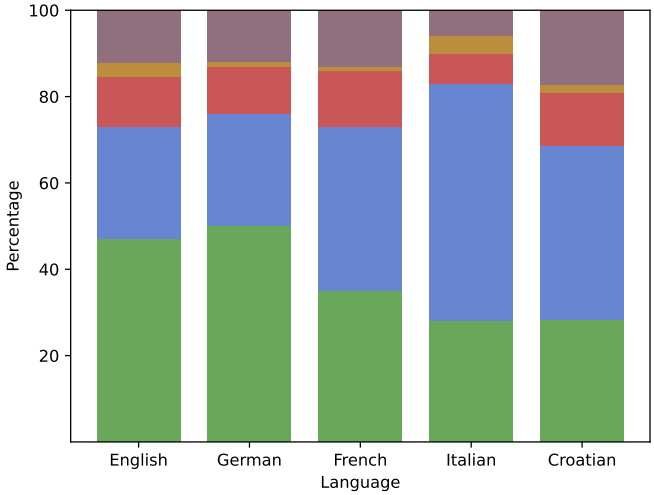


Figure 4: Evaluation of found terms.

Table 4: Evaluation of error causes in false positives.

Error type	Percentage
Wrong document	56.77%
Technical issue	24.52%
Unusual website design	12.26%
Incomplete document	6.45%

5 ACCESS

To foster empirical research of policies and terms, we are open to providing access to the database for performing large-scale studies. Please indicate your interest on the project page at the following link https://karelkubicek.github.io/post/pptc. We have published the anonymized data (see Section 7) from twelve months of scraping in 2024 on Zenodo (https://doi.org/10.5281/zenodo.14562039).

6 NEXT STEPS AND FUTURE WORK

We have presented and implemented a system for the automated identification and storage of policies and terms on websites on a large scale (800 000 websites). As described in Section 4, our scraper does a reasonable job for major languages such as English, German, French, or Italian. But its performance for less common languages such as Croatian is more limited. We cannot exclude that such lower performance is a trend that spreads across all less common languages. We view the scraper as a dynamic project that will benefit from further research by our team and hopefully other interested researchers as well.

For applied research, we anticipate that the policies and terms in our database will serve as a basis for empirical analyses in various fields. We hope that our scraper will contribute to unifying the study of policies and terms, and improve the comparability and reproducibility of research findings.

Potential research ideas using our database could entail, but are not limited to:

(1) *Evolution over time*: Researchers can study how policies and terms change over time, be it through endogenous innovation or through exogenous shocks such as through new regulations.
(2) *Linguistic heterogeneity*: Until very recently, studies of policies and terms were mostly limited to the English language. With our dataset, scholars can inspect whether previous and novel findings are robust to other languages.
(3) *Scale and representative sample*: Our large and representative sample enables to study policies and terms on the internet holistically without a bias for more popular websites. As the CrUX list contains popularity buckets for individual websites, researchers can, for example, specifically study trends for less popular websites.
(4) *Interaction of policies and terms*: The unique linkage of policies and terms of individual websites offered by our dataset opens up a wide range of research opportunities.

To foster empirical research of policies and terms, we are currently working on providing access to the database for parties interested in performing large-scale studies. We are also working on creating an intuitive interface to visualize changes of policies and terms over time.

7 CONCLUSION

While policies and terms have been analyzed by researchers from law, computer science, and economics for some time, no large database of multilingual policies and terms with long-term support exists. Creating new corpora of policies and terms for individual research projects is not only inefficient, but it also hampers the reproducibility and comparability of studies. To fill this gap, we introduce a scraper for the automated identification and storage of policies and terms on hundreds of thousands of websites. Our results show F1 scores of 79% for policies and 75% for terms, with even better performance for more common languages. We introduce several broad research ideas using our database and potential refinements of our scraper as future work.

ETHICAL CONSIDERATIONS

We are aware that the creation and publication of our dataset may entail risks related to privacy and website overload. To address these risks, we implemented the following precautionary measures:

(1) We evaluated all 155 false positive (see Section 4) documents for any privacy related or other risks. We found 12 documents (7.74%) containing personal data. Therefore, we anonymized our entire published dataset with the tool Presidio [35]. The only other risk we encountered was a forum with potential vaccine-related misinformation. However, we considered this an acceptable risk, as the scraper's purpose is not to provide accurate information on scientific topics.
(2) We do not load websites extensively, as we make at most 10 requests to the website per month. This procedure does not result in noticeable costs or service limitations for real users.
(3) We only collect publicly available website content.
(4) We do not create any content on the website.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the works of Luka Lodrant, who implemented key concepts of the scraper as part of his master thesis and Truong Hoang Long, who adapted the scraper for its current purpose of searching for policies and terms as part of his bachelor thesis. Elias Landes and Natalie Ashgriz provided excellent research assistance. We also thank Joachim Posegga from the University of Passau and the German Research Network (DFN) for providing us with a VPN. Karel Kubicek's research is funded by SNSF Postdoc.Mobility grant No. P500PT_225449. Luka Nenadic's research is funded by SNSF grant No. 10002634.

REFERENCES

[1] Andrick Adhikari, Sanchari Das, and Rinku Dewri. 2023. Evolution of composition, readability, and structure of privacy policies over two decades. *Proceedings on Privacy Enhancing Technologies* 2023, 3 (2023), 138–153. https://doi.org/10.56553/popets-2023-0074
[2] 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, SI) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 2165–2176. https://doi.org/10.1145/3442381.3450048
[3] 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 Rrgimes: Creation and analysis of a bilingual privacy policy corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, FR, 5460–5472. https://aclanthology.org/2022.lrec-1.585
[4] 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.
[5] Shmuel I. Becher and Uri Benoliel. 2021. Law in books and law in action: The readability of privacy policies and the GDPR. In *Consumer Law and Economics*,

Klaus Mathis and Avishalom Tor (Eds.), Vol. 9. Springer, Cham, CH, 179–204. https://doi.org/10.1007/978-3-030-49028-7_9

[6] Uri Benoliel and Shmuel I. Becher. 2024. Messy contracts. *University of Illinois Law Review* 3 (2024), 893–938. https://illinoislawreview.org/wp-content/uploads/2024/07/Benoliel.pdf

[7] Francesco Ciclosi, Silvia Vidor, and Fabio Massacci. 2023. Building cross-language corpora for human understanding of privacy policies. In *Digital Sovereignty in Cyber Security: New Challenges in Future Vision*, Antonio Skarmeta, Daniele Canavese, Antonio Lioy, and Sara Matheu (Eds.). Springer Nature Switzerland, Cham, CH, 113–131. https://doi.org/10.1007/978-3-031-36096-1_8

[8] 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

[9] Kevin E Davis and Florencia Marotta-Wurgler. 2019. Contracting for personal data. *New York University Law Review* 94 (2019), 662–705. https://www.nyulawreview.org/issues/volume-94-number-4/contracting-for-personal-data/

[10] Kevin E. Davis and Florencia Marotta-Wurgler. 2024. Filling the void: How E.U. privacy law spills over to the U.S. *Journal of Law and Empirical Analysis* 1, 1 (2024), 1–21. https://doi.org/10.1177/2755323X241237619

[11] 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, 1–15. https://doi.org/10.14722/ndss.2019.23378

[12] Nurullah Demir, Matteo Große-Kampmann, Tobias Urban, Christian Wressnegger, Thorsten Holz, and Norbert Pohlmann. 2022. Reproducibility and replicability of web measurement studies. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, FR) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 533–544. https://doi.org/10.1145/3485447.3512214

[13] Ricardo Dominguez-Olmedo, Vedant Nanda, Rediet Abebe, Stefan Bechtold, Christoph Engel, Jens Frankenreiter, Krishna Gummadi, Moritz Hardt, and Michael Livermore. 2024. Lawma: The power of specialization for legal tasks. https://doi.org/10.48550/arXiv.2407.16615 arXiv:2407.16615 [cs].

[14] Kasper Drawzeski, Andrea Galassi, Agnieszka Jablonowska, Francesca Lagioia, Marco Lippi, Hans Wolfgang Micklitz, Giovanni Sartor, Giacomo Tagiuri, and Paolo Torroni. 2021. A corpus for multilingual analysis of online terms of service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, Nikolaos Aletras, Ion Androutsopoulos, Leslie Barrett, Catalina Goanta, and Daniel Preotiuc-Pietro (Eds.). Association for Computational Linguistics, Punta Cana, DO, 1–8. https://doi.org/10.18653/v1/2021.nllp-1.1

[15] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. 2017. Large-scale readability analysis of privacy policies. In *Proceedings of the International Conference on Web Intelligence*. ACM, Leipzig, DE, 18–25. https://doi.org/10.1145/3106426.3106427

[16] Jens Frankenreiter. 2022. Cost-based California Effects. *Yale Journal on Regulation* 39 (2022), 1155. https://www.yalejreg.com/print/cost-based-california-effects/

[17] Andrea Galassi, Kasper Drazewski, Marco Lippi, and Paolo Torroni. 2020. Cross-lingual annotation projection in legal texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 915–926. https://doi.org/10.18653/v1/2020.coling-main.79

[18] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. https://digitalcommons.osgoode.yorku.ca/all_papers/380

[19] 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

[20] 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

[21] 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

[22] Noam Kolt. 2022. Predicting consumer contracts. *Berkeley Technology Law Journal* 37, 1 (2022), 71–138. https://doi.org/10.15779/Z382B8VC90

[23] 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

[24] F. Lagioia, A. Jabłonowska, R. Liepina, and K. Drazewski. 2022. AI in search of unfairness in consumer contracts : the terms of service landscape. *Journal of Consumer Policy* 45, 3 (Sept. 2022), 481–536. https://doi.org/10.1007/s10603-022-09520-9

[25] Logan Lebanoff and Fei Liu. 2018. Automatic detection of vague words and sentences in privacy policies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, BE, 3508–3517. https://doi.org/10.18653/v1/D18-1387

[26] R. Liepiņa, Giuseppe Contissa, K. Drazewski, F. Lagioia, Marco Lippi, H. Micklitz, Przemyslaw Palka, G. Sartor, and Paolo Torroni. 2019. GDPR privacy policies in CLAUDETTE: Challenges of omission, context and multilingualism. In *Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Text*. RWTH Aachen, Montreal, CA, 1–7. https://www.semanticscholar.org/paper/GDPR-Privacy-Policies-in-CLAUDETTE%3A-Challenges-of-Liepi%C5%86a-Contissa/fc2daaa4171db92bd30cc67199374cda06b562ca

[27] 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

[28] Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Yannis Panagis, Giovanni Sartor, and Paolo Torroni. 2017. Automated detection of unfair clauses in online consumer contracts. In *Legal Knowledge and Information Systems*. IOS Press BV, Amsterdam, NL, 145–154. https://doi.org/10.3233/978-1-61499-838-9-145

[29] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law* 27, 2 (June 2019), 117–139. https://doi.org/10.1007/s10506-019-09243-2

[30] Souvik Maitra and Dwijen Rudrapal. 2023. Comparative study on different approaches for understanding the privacy policies. In *Evolution in Computational Intelligence*, Vikrant Bhateja, Xin-She Yang, Jerry Chun-Wei Lin, and Ranjita Das (Eds.). Springer Nature, Singapore, 15–23. https://doi.org/10.1007/978-981-19-7513-4_2

[31] Florencia Marotta-Wurgler. 2016. Self-regulation and competition in privacy policies. *Journal of Legal Studies* 45 (2016), 13–39.

[32] Florencia Marotta-Wurgler and Robert Taylor. 2013. Set in stone? Change and innovation in consumer standard-form contracts. *New York University Law Review* 88 (2013), 240–285. https://www.nyulawreview.org/wp-content/uploads/2018/08/NYULawReview-88-1-MarottaWurgler-Taylor.pdf

[33] 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

[34] Hans-W. Micklitz, Przemysław Pałka, and Yannis Panagis. 2017. The empire strikes back : Digital control of unfair terms of online services. *Journal of Consumer Policy* 40, 3 (Sept. 2017), 367–388. https://doi.org/10.1007/s10603-017-9353-0

[35] Microsoft. 2025. Presidio: Data Protection and De-identification SDK. https://github.com/microsoft/presidio/ Accessed: 2025-01-14.

[36] Jayashree Mohan, Melissa Wasserman, and Vijay Chidambaram. 2019. Analyzing GDPR compliance through the lens of privacy policy. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, Vijay Gadepally, Timothy Mattson, Michael Stonebraker, Fusheng Wang, Gang Luo, Yanhui Laing, and Alevtina Dubovitskaya (Eds.). Springer International Publishing, Cham, 82–95.

[37] Huss Nick. 2024. How many websites are there in the world? (2024). https://siteefy.com/how-many-websites-are-there/

[38] Jonathan A. Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147. https://doi.org/10.1080/1369118X.2018.1486870

[39] Aaron Perzanowski and Jason Schultz. 2016. *The End of Ownership: Personal Property in the Digital Economy*. MIT Press, Cambridge, MA.

[40] 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

[41] 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, 374–387. https://doi.org/10.1145/3517745.3561444

[42] 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, et al. 2013. Usable privacy policy project. In *Technical report, Technical Report, CMU-ISR-13-119*. Carnegie Mellon University, Pittsburgh, Pennsylvania. https://usableprivacy.org/data

[43] Tim Samples, Katherine Ireland, and Caroline Kraczon. 2024. TL;DR: The law and linguistics of social platform terms-of-use. *Berkeley Technology Law Journal* 39 (2024), 47–108. https://doi.org/10.15779/Z38HT2GC9C
[44] W. David Slawson. 1971. Standard form contracts and democratic control of lawmaking power. *Harvard Law Review* 84, 3 (1971), 529–566. http://www.jstor.org/stable/1339552
[45] Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6829–6839. https://doi.org/10.18653/v1/2021.acl-long.532
[46] Van Hong Tran, Aarushi Mehrotra, Marshini Chetty, Nick Feamster, Jens Frankenreiter, and Lior Jacob Strahilevitz. 2024. Measuring compliance with the California Consumer Privacy Act over space and time. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, May 11-16, 2024*, Florian 'Floyd' Mueller, Penny Kyburz, Julie R. Williamson, Corina Sas, Max L. Wilson, Phoebe O. Toups Dugas, and Irina Shklovski (Eds.). ACM, Honolulu, HI, 785:1–785:19. https://doi.org/10.1145/3613904.3642597
[47] Isabel Wagner. 2023. Privacy policies across the ages: Content of privacy policies 1996–2021. *ACM Transactions on Privacy and Security* 26, 3 (2023), 1–32.
[48] 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, DE, 1330–1340. https://doi.org/10.18653/v1/P16-1126
[49] Razieh Nokhbeh Zaeem and K. Suzanne Barber. 2021. The effect of the GDPR on privacy policies: Recent progress and future promise. *ACM Transactions on Management Information Systems* 12, 1 (March 2021), 1–20. https://doi.org/10.1145/3389685
[50] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R. Reidenberg, N. Cameron Russell, and Norman M. Sadeh. 2019. MAPS: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies* 2019 (2019), 66 – 86.

A APPENDIX

Supported Languages and Countries

Supported languages. Our scraper supports 37 languages, with most keywords translated by native or proficient speakers who observed multiple websites prior to the translation.[7] These languages are: Albanian*, Basque*, Bosnian, Bulgarian, Catalan*, Croatian, Czech, Danish, Dutch, English, Estonian*, Finnish*, French, Galician, German, Greek*, Hungarian*, Icelandic*, Italian, Latvian*, Lithuanian*, *Luxembourgish**, Macedonian*, *Maltese**, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Turkish*, Ukrainian, and Welsh*.

Luxembourgish and Maltese (italicized) are not supported by multilingual BERT, which we use for document classification, and, therefore, have limited support.

Selected countries and maximal ranks. We sample CrUX for the following list of countries:[8] Austria (500k), Belgium (500k), Bulgaria (500k), Croatia (100k), Cyprus (50k), Czechia (500k), Denmark (500k), Estonia (100k), Finland (500k), France (1M), Germany (1M), Great Britain (5M), Greece (500k), Hungary (500k), Iceland (50k), Ireland (500k), Italy (1M), Latvia (100k), Lichtenstein (50k), Lithuania (100k), Luxembourg (100k), Malta (50k), Netherlands (1M), Norway (500k), Poland (1M), Portugal (500k), Romania (500k), Russia (1M), Slovakia (500k), Slovenia (500k), Spain (1M), Sweden (500k), Switzerland (500k), Turkey (1M), and the United States (5M).

[7] A star was added to languages were keywords were translated by Google Translate.
[8] Each country has a maximal rank in which there are websites available. We denote the rank in parentheses.

Bot-Evasion Techniques

To increase representativeness of scraped policies and terms, we use the following techniques to decrease the likelihood of our scraper being detected as a bot.

Browser. The scraper uses a full Chrome browser, namely the implementation of Undetected Chromedriver, which extends the Selenium Chromedriver by removing Selenium fingerprints, and numerous other bot-evasion techniques.

Browser fingerprint. To interpret the load status of each visited page, we use Chrome DevTools Protocol (*CDP*). Unlike using a proxy to intercept the traffic, CDP is not prone to TLS fingerprinting, where the proxy and the browser ciphersuits mismatch. Such a mismatch is then detectable by modern bot detection systems such as Cloudflare. We also run Chrome with a non-root user, else Chrome disables sandboxing protections, which is again detectable by services such as Cloudflare.

VPN. The ability to select a vantage point is important for many studies of policies and terms. However, common commercial VPN endpoints are publicly known and therefore often blocked. Demir et al. [12] observed that residential proxies are the least likely to correspond to bot activity, closely followed by university VPNs, while datacenters and commercial VPNs are blocked more frequently. For our mostly EU sample, we used six IP addresses provided by a German university.

Codebook for Annotators

This section summarizes the codebook (see Section 4) used to annotate policies and terms. To find the policy or terms of a given website, for the vast majority of websites, the relevant document could be retrieved at the bottom of the page by clicking on the corresponding link. If there was no such link, annotators would search for the relevant document using a search engine. The annotators then annotated the relevant website along the following categories:

(1) The ground truth URL (*true URL*) for the policy or terms;
(2) Whether the true URL matches with the scraper's URL;
(3) Whether the content of the true URL matches the scraper's identified content by looking at the first paragraph;
(4) The error type (see Section 4);
(5) Comments for special cases.

Neither cookie notices nor sections of terms related to privacy were considered to be policies. Terms were only considered if they were general terms, not if they were special terms for specific services or specific parts of the website. If there were general terms, these alone were considered as terms even if additional specific terms (e.g., terms for suppliers) existed. We acknowledge that this differentiation may not always be crystal clear and that certain gray areas could not be avoided. For example, "terms of sale" may be the general terms on an e-commerce website, but only be specific terms in addition to general terms on other types of websites.

Mere legal mentions, even if they contained more than just the name of the site developer and/or hosting provider, were not considered to be terms or policies. Exceptions were only made where a "full" policy or "full" terms could be found therein.