Privacy Now or Never: Large-Scale Extraction and Analysis of Dates in Privacy Policy Text

Mukund Srinath
mukund@psu.edu
Pennsylvania State University
University Park, PA, USA

Lee Matheson
lmatheson@fpf.org
Future of Privacy Forum
Washington DC, USA

Pranav Narayanan Venkit
pranav.venkit@psu.edu
Pennsylvania State University
University Park, PA, USA

Gabriela Zanfir-Fortuna
gzanfir-fortuna@fpf.org
Future of Privacy Forum
Washington DC, USA

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

C. Lee Giles
clg20@psu.edu
Pennsylvania State University
University Park, PA, USA

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

ABSTRACT

The General Data Protection Regulation (GDPR) and other recent privacy laws require organizations to post their privacy policies, and place specific expectations on organisations' privacy practices. Privacy policies take the form of documents written in natural language, and one of the expectations placed upon them is that they remain up to date. To investigate legal compliance with this recency requirement at a large scale, we create a novel pipeline that includes crawling, regex-based extraction, candidate date classification and date object creation to extract updated and effective dates from privacy policies written in English. We then analyze patterns in policy dates using four web crawls and find that only about 40% of privacy policies online contain a date, thereby making it difficult to assess their regulatory compliance. We also find that updates in privacy policies are temporally concentrated around passage of laws regulating digital privacy (such as the GDPR), and that more popular domains are more likely to have policy dates as well as more likely to update their policies regularly.

CCS CONCEPTS

• **Information systems → Data extraction and integration**.

KEYWORDS

privacy policy, date extraction, crawling

ACM Reference Format:
Mukund Srinath, Lee Matheson, Pranav Narayanan Venkit, Gabriela Zanfir-Fortuna, Florian Schaub, C. Lee Giles, and Shomir Wilson. 2023. Privacy Now or Never: Large-Scale Extraction and Analysis of Dates in Privacy Policy Text. In *ACM Symposium on Document Engineering 2023 (DocEng '23), August 22–25, 2023, Limerick, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3573128.3609342

1 INTRODUCTION

Privacy policies are legal documents that organizations use to disclose how they collect, analyze, share, and secure their users' personal data. Various laws around the world, such as the GDPR (General Data Protection Regulation), CCPA (California Consumer Privacy Act) amended to CPRA (California Privacy Rights Act) and VCDPA (Virginia Consumer Data Protection Act), place specific requirements on the content of privacy policies and make it mandatory for organizations to make their policies publicly available. More than 182 legal jurisdictions around the world have some form of privacy laws, and globally, the number of privacy laws has been growing exponentially since the 21st century [8].

Most privacy regulations rely on the *notice and choice* framework, where organizations are required to give their users a *notice* of their data practices, and users then have the *choice* to either use the services offered or decline to proceed [14]. The implicit assumption is that users will read and understand organizations' policies and that organizations will keep their policies up-to-date and accessible to users [3]. However, users might not always be able to read or understand policies, due to various failure modes such as broken policy links, natural language discrepancies, policies being too long and complicated, and empty content [18].

In this paper, we investigate one potential failure mode in privacy policy availability, namely whether the policy is up-to-date, and study the effects of laws and regulations on policy updates. Extracting dates from policies is crucial for informing trends in the online privacy landscape and help study historical user privacy cause-and-effect relationships. Additionally, systematically extracting dates at scale from policies will aid in privacy regulation enforcement. The GDPR (Articles 12-14), explicitly requires transparent disclosure of information, accurately reflecting data practices. The lack of a policy date would create doubt about whether the policy is reflecting

DocEng '23, August 22–25, 2023, Limerick, Ireland

ACM ISBN 979-8-4007-0027-9/23/08…$15.00
https://doi.org/10.1145/3573128.3609342

up-to-date data practices therefore violating this requirement in principle. The privacy policy is the only document that consumers can use to understand what happens with their personal information, and without a policy date, a consumer would not be able to ascertain if a policy is current or up-to-date. Additionally, the lack of posted policy dates hampers regulatory oversight, by making it difficult or impossible to determine when a policy was effective.

We use four large web crawls, containing over 3.5 million privacy policies, to study trends in dates and updates in privacy policies online. We create a pipeline to extract policy dates by first extracting text from the privacy policy HTML, followed by extracting all date instances and classifying candidate policy dates into either policy date, i.e., *updated date* or *effective date*, or other types of dates using a transformer-based language model[1]. We find that less than 40% of policies online have a policy date and that popular domains are more likely to have a policy date. Finally, we investigate trends in updating privacy policies and find that the dates when new privacy laws became effective have had a significant impact on when most privacy policies have been updated. Our findings strongly suggest that a large percentage of policies online are not up-to-date and are not compliant with legal requirements.

2 RELATED WORK

Research on extracting dates from unstructured documents has focused on the news domain, and used named entity recognition (NER) approaches. TimeLineCurator extracts all dates in news articles using a supervised NER approach [7]. News articles are often dense in dates since they recount events. NER is thus the most appropriate technique since temporal features could be recounted in various ways, such as the day of the week, day of the month, etc. Similarly, Smith [15] finds date-place co-locations in news articles.

Work in unstructured date extraction, similar to that in this paper, can be found in the medical domain. Fu et al. [6] extract diagnosis dates from clinical notes using a regex approach followed by a machine learning model to classify the extracted dates. However, they test their approach on a very small document set with dates occurring in similar formats. Similarly, studies on extracting clinical information from patient health records often also include extracting a date [5, 13]. However, they often contain much richer context, thereby helping to classify date types.

Prior research studying patterns in privacy policies' dates is scarce. Linden et al. [10] and Degeling et al. [4] studied how GDPR changed the privacy landscape using the Wayback Machine to download policies before and after the GDPR and compared changes between them. PrivaSeer, the privacy policy search engine [16] displays the date that a policy was previously crawled and makes them searchable. However, no prior work concentrates on extracting dates from policies in order to analyze policy update patterns.

3 POLICY DATE EXTRACTION

We attempt to solve the problem of extracting the updated and effective dates given a large corpus of privacy policies while optimizing for time and computational resources. We create a policy date extraction pipeline [17, 19] in which we first extract text snippets containing candidate policy dates using a regex-based approach. We then classify the candidate instances and construct date objects using an open-source date extraction tool. We then analyze trends in dates in over 3.5 million web privacy policies using four large-scale web crawls between May 2019 and September 2021.

Regex Extraction: We surveyed a number of tools that automatically extract dates from unstructured text. Table 1 shows the tools and how they perform on a random sample of one hundred documents. The tools surveyed, namely, Spacy [9], SUTtime [1], dateparser[2] and datefinder[3] extract too many non-policy date instances with dateparser and datefinder extracting almost 22 instances per policy, thereby requiring more processing further down the pipeline. SUTime takes almost 5s per policy on a single CPU core, translating to about half a year to extract date instances from 3.5 million policies. Although Spacy's named entity recognition (NER) approach achieves reasonable results, our custom approach achieves a multi-fold improvement on time and precision. Our custom regex concentrates on matching common date patterns in privacy policies, such as, four-digit numbers starting with 20 or 19, (any policy would need to be written between the late 1990s and today) and a combination of separators, date, month and year. We then extract at most 250 characters (or up to a newline character) of text before and after the regex match to capture the context of the matched four-digit number.

Table 1: Performance of open-source date extraction tools (P: Precision; R: Recall; time (seconds) indicated per policy; instances (mean) per policy)

Tool	Time (s)	Instances	P	R
datefinder	0.0005	21.8	0.01	0.75
dateparser	0.1	21.8	0.01	1
SUTime	5	5.5	0.04	1
Spacy (NER)	0.24	4.6	0.05	1
Custom Regex	**0.0005**	**3.0**	**0.13**	**1**

Date Instance Classification: Privacy policies contain a number of date instances. Most commonly, they contain when the policy was updated, when the policy went into effect, date of previous revision, version approval date, and regulation enforcement date. Since we were interested in studying the trends of updating privacy policies, we concentrated on extracting *updated* and *effective* dates. We refer to these as 'policy dates'. The following are a few examples of types of policy date instances we extract: *Effective: January 1st, 2022*, *This policy went into effect on 1/1/22*, *The policy dated January 1, 2022, replaces all previous policies*, *Last updated: January 2022*, *Last Revised 01/01/2022*, *This page was last edited on 1 Jan 2022*.

We use a supervised classification approach to separate policy dates from all the date instances extracted. One of the authors, a privacy policy expert, labeled 1,075 uniformly randomly sampled date instances. 230 were found to be *updated dates*, 170 were *effective dates* and the rest (675) were *other* date instances. We trained two models to classify the extracted date instances. We trained a random-forest model with 100 estimators and a minimum of 2 samples to split a node on tf-idf features. We also fine-tuned PrivBERT [17],

[1] Data and code available at https://privaseer.ist.psu.edu/data

[2] https://github.com/scrapinghub/dateparser

[3] https://github.com/akoumjian/datefinder

a language model pretrained on privacy policies, using the Adam optimizer with a training rate of 1e-4 and batch size of 64 for three epochs. We divided the labeled documents into train, test, and validation sets in the ratio 3:1:1. Table 2 shows the results for this task. We can see that PrivBERT improves the performance of the baseline approach by a few percentage points.

Table 2: Date instance classification results (S: Support)

Class	Random Forest			PrivBERT			S
	P	R	F1	P	R	F1	
Updated	0.98	0.92	0.95	0.95	0.97	**0.96**	48
Effective	0.95	0.92	0.93	0.97	0.97	**0.97**	38
Other	0.97	1.0	0.98	0.99	0.99	**0.99**	129

Date Object Extraction: After the date instances were classified as either a policy date or not, we used dateparser to create date objects, i.e., separately representing date text instances into day, month, and year. A number of tools exist that convert date instances into objects, some of which we surveyed in the section on regex extraction shown in Table 1. We used dateparser since it was quick, had a high accuracy rate, and was the most convenient.

4 POLICY DATE ANALYSIS

We created a pipeline to efficiently collect privacy policies at scale and used four large-scale web crawls to study date patterns in them. Our pipeline involved first collecting candidate privacy policy URLs from the CommomCrawl URL dump using a selection criterion: URLs containing the words 'privacy' or the words 'data' and 'protection.' We then crawled the URLs satisfying this criterion and filtered the English language documents using LangID [11]. Next, we trained a model to classify whether a given document is a privacy policy [17]. We obtained a precision of 0.98, and a recall of 0.98 [17]. Next, we removed duplicates by hashing the HTML pages, and removed near-duplicates by creating simhashes (similar items have similar hashes) [2] and excluding documents based on a Hamming distance threshold [12].

We undertook the first crawl in May 2019 and obtained about 1 million privacy policies [17]. We completed the second crawl in March 2020, starting with a LinkedIn company dataset, and obtained about 600k policies [18]. We completed the third crawl in March 2021, starting with a more recent CommonCrawl archive, and created a collection of about 1.2 million policies. Finally, we undertook the fourth crawl in September 2021, when we re-crawled all the URLs in the first, and second crawls. Overall, we collected over 3.5 million policies, including multiple versions of the same policy.

We put all four crawls through the date extraction pipeline. We found that only a small portion of the policies in any crawl had a policy date, with the March 2020 crawl containing the highest percentage (41.6%) compared to 36.6% and 38% in the May 2019 and March 2021 crawls respectively. This is likely due to the source of the seed URLs used. The seed URLs for the March 2019 crawl came from the company URLs listed on LinkedIn, whereas CommonCrawl was the source used for the other crawls. CommonCrawl seeds are more likely to contain links to hobby sites, information sites, and other non-commercial sites that are less likely to maintain up-to-date privacy policies and follow regulatory updates. More evidence for this hypothesis can be found in Figure 1, which shows how the presence of policy dates varies based on the pagerank of the domain as calculated from the CommonCrawl web-graph. In the figure, the x-axis contains bins of pagerank values, where the higher the pagerank value, the more popular the domain. Each bin contains at least 50 policies, with the number of policies skewed heavily towards the left (less popular domains). We can see that more popular domains are more likely to contain policy date instances. This is likely due to the fact that more popular sites face heavier regulatory scrutiny.

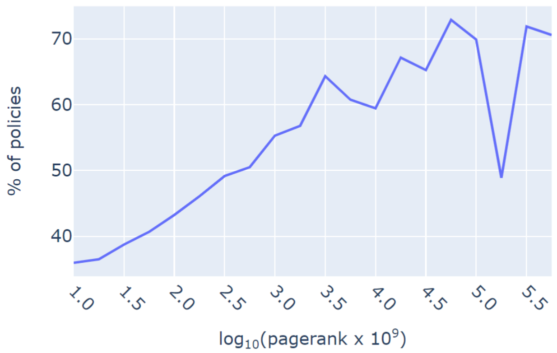

Figure 1: Relationship between %policies containing at least one date instance and the domain pagerank value

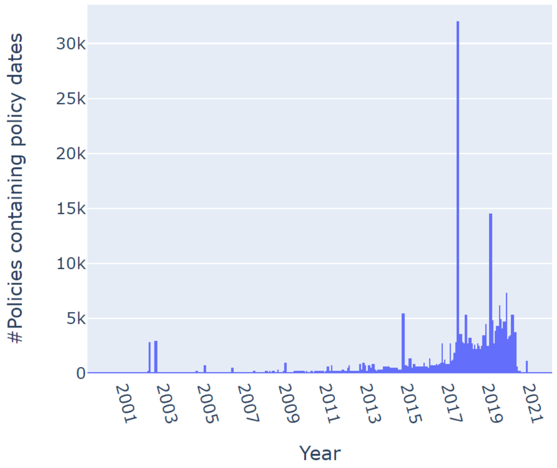

Figure 2: Policy date distribution

Figure 2 shows that a large percentage of policies have a policy date in May 2018. This corresponded with when GDPR came into effect, i.e., May 25, 2018. Degeling et al. [4] noted a similar increase in updates to privacy policies leading up to and around the GDPR effective date. This shows that many policies updated to comply with GDPR have not been updated in the four years since. Further, the second most significant spike in the graph occurred in December 2019, corresponding with the CCPA coming into effect starting 1

January 2020. The earlier peak around April 2003, corresponds with the passage of the Privacy Act of 2003. These peaks in the policy update patterns surrounding regulation enforcement dates suggest that organizations updated their policies with the passage of new regulations but failed to keep their policies regularly updated.

We investigated how often policies were updated by measuring the number of overlapping policies between each crawl and then counting the number of policies whose policy dates changed. Between the May 2019 and September 2021 crawls, i.e., in over two years, we found that only about 32% of policies were updated. Between the May 2019 and March 2020 crawls, 27% of policies were updated, while between the March 2020 and March 2021 crawls, 29% of policies were updated, respectively.

Finally, we investigated the likelihood of policies having a date based on their sector of commerce. [18] categorized the industry information (obtained through LinkedIn) for all the domains in the March 2020 crawl into ten sectors of commerce. We found that the *Information Technology* sector and the *Education* sector has the highest percentage of policies, with 42.24% and 42.1% of policies having a policy date respectively, while *Consumer and supply chain* sector is the one with the lowest percentage of policies, with around 39.7%, containing a policy date. Overall, we hypothesize that sectors that are more likely to be consumer-facing have a higher likelihood of having policy dates.

5 DISCUSSION AND CONCLUSION

The privacy policy is the only document consumers can use to understand what happens with their personal information online. It is, therefore, paramount that organizations reflect up-to-date information regarding their data practices on their policies. Figure 2 shows how a number of policies have been dormant for up to 20 years (2000-2020). In contrast, real-world data collection practices have undergone a multitude of significant changes during the same period, suggesting that these policies may not reflect up-to-date data collection practices of their respective organizations. A lack of policy dates can thus create confusion regarding whether a policy is being updated or maintained from a user's point of view and thus disincentivizing them to read policies, and further contributing to the failures of the notice and choice model. Our findings shown in Figure 1 suggest that even among popular domains (with a high pagerank), only about 70% contain a policy date with fewer updating their policies annually. This percentage falls to less than 40% of policies when considering less popular domains. Additionally, our findings show that at scale, this trend has held steady between May 2019 and September 2021 suggesting that a more stringent regulatory incentive is required for organisations to include a policy dates in their documents and keep them updated.

In this paper, we created a novel pipeline to extract updated and effective dates from privacy policies at scale while optimizing for time and computational resources. The pipeline includes crawling, regex extraction, date instance classification, and date object extraction, and can be applied to extract date instances at scale for any domain by retraining the machine learning model. We crawled and applied the pipeline on over 3.5 million privacy policies and extracted policy dates from them. We found that less than 40% of policies have at least a single policy date and less than 30% of which are updated annually. Further, we found that policies tend to be updated en mass as a consequence of new regulation enforcement and that more popular domains are more likely to update policies or even contain a policy date. Finally, we found that the consumer-facing sectors had the highest percentage of policies with dates, with about 42% containing a policy date.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Angel X Chang and Christopher D Manning. 2012. Sutime: A library for recognizing and normalizing time expressions.. In *Lrec*, Vol. 3735. 3740.

[2] Moses S Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. ACM, 380–388.

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

[4] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2018. We value your privacy... now take some cookies: Measuring the GDPR's impact on web privacy. *arXiv preprint arXiv:1808.05096* (2018).

[5] Beata Fonferko-Shadrach, Arron S Lacey, Angus Roberts, Ashley Akbari, Simon Thompson, David V Ford, Ronan A Lyons, Mark I Rees, and William Owen Pickrell. 2019. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system. *BMJ open* 9, 4 (2019), e023232.

[6] Julia T Fu, Evan Sholle, Spencer Krichevsky, Joseph Scandura, and Thomas R Campion. 2020. Extracting and classifying diagnosis dates from clinical notes: a case study. *Journal of Biomedical Informatics* 110 (2020), 103569.

[7] Johanna Fulda, Matthew Brehmer, and Tamara Munzner. 2015. TimeLineCurator: Interactive authoring of visual timelines from unstructured text. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 300–309.

[8] Sonu Gupta, Ellen Poplavska, Nora O'Toole, Siddhant Arora, Thomas Norton, Norman Sadeh, and Shomir Wilson. 2022. Creation and Analysis of an International Corpus of Privacy Laws. *arXiv preprint arXiv:2206.14169* (2022).

[9] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. (2020).

[10] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2020. The Privacy Policy Landscape After the GDPR. *Proceedings on Privacy Enhancing Technologies* 1 (2020), 47–64.

[11] Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*. Association for Computational Linguistics, 25–30.

[12] Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 141–150.

[13] Anoop D Shah, Carlos Martinez, and Harry Hemingway. 2012. The freetext matching algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records. *BMC medical informatics and decision making* 12, 1 (2012), 1–13.

[14] Robert H Sloan and Richard Warner. 2014. Beyond notice and choice: Privacy, norms, and consent. *J. High Tech. L.* 14 (2014), 370.

[15] David A Smith. 2002. Detecting events with date and place information in unstructured text. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*. 191–196.

[16] Mukund Srinath, Soundarya Nurani Sundareswara, C Lee Giles, and Shomir Wilson. 2021. PrivaSeer: A Privacy Policy Search Engine. In *International Conference on Web Engineering*. Springer, 286–301.

[17] 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 (Volume 1: Long Papers)*. 6829–6839.

[18] Soundarya Sundareswara, Shomir Wilson, Mukund Srinath, and Lee Giles. 2020. Privacy not found: a study of the availability of privacy policies on the web.

[19] Soundarya Nurani Sundareswara, Mukund Srinath, Shomir Wilson, and C. Lee Giles. 2021. A Large-Scale Exploration of Terms of Service Documents on the Web. In *Proceedings of the 21st ACM Symposium on Document Engineering* (Limerick, Ireland) *(DocEng '21)*. Association for Computing Machinery, New York, NY, USA, Article 21, 4 pages. https://doi.org/10.1145/3469096.3474940