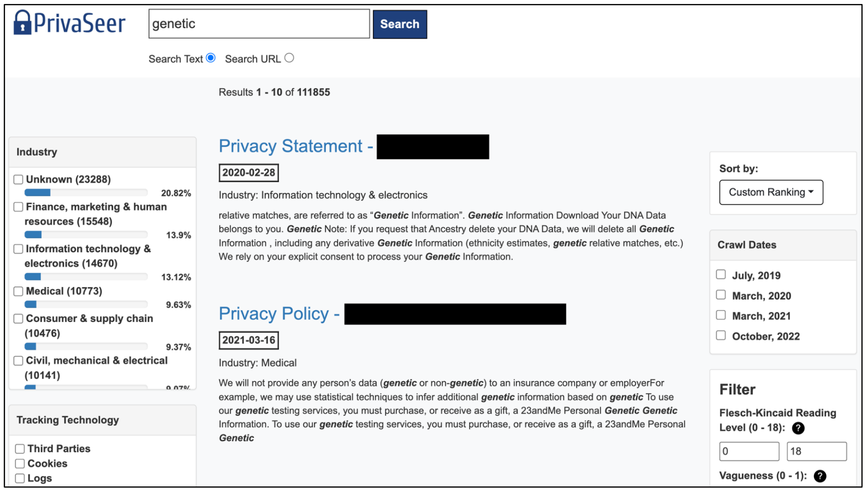
The PrivaSeer Project: Large-Scale Resources for Analysis of Privacy Policy Text

Shomir Wilson[1], Florian Schaub[2], Lee Matheson[3], Shahriar Shayesteh[1], and Lu Xian[2]
[1]College of Information Sciences and Technology, Pennsylvania State University
[2]School of Information, University of Michigan, [3]Future of Privacy Forum

You can download the PrivaSeer Corpus of 3,967,487 privacy policies and use the PrivaSeer Search Engine at https://privaseer.ist.psu.edu/


PrivaSeer

genetic | Search

Search Text ● Search URL ○

Results **1 - 10** of **111855**

Industry

- ☐ Unknown (23288) 20.82%
- ☐ Finance, marketing & human resources (15548) 13.9%
- ☐ Information technology & electronics (14670) 13.12%
- ☐ Medical (10773) 9.63%
- ☐ Consumer & supply chain (10476) 9.37%
- ☐ Civil, mechanical & electrical (10141) 9.07%

Tracking Technology

- ☐ Third Parties
- ☐ Cookies
- ☐ Logs

Privacy Statement - [redacted]

2020-02-28

Industry: Information technology & electronics

relative matches, are referred to as "***Genetic*** Information". ***Genetic*** Information Download Your DNA Data belongs to you. ***Genetic*** Note: If you request that Ancestry delete your DNA Data, we will delete all ***Genetic*** Information , including any derivative ***Genetic*** Information (ethnicity estimates, ***genetic*** relative matches, etc.) We rely on your explicit consent to process your ***Genetic*** Information.

Privacy Policy - [redacted]

2021-03-16

Industry: Medical

We will not provide any person's data (***genetic*** or non-***genetic***) to an insurance company or employerFor example, we may use statistical techniques to infer additional ***genetic*** information based on ***genetic*** To use our ***genetic*** testing services, you must purchase, or receive as a gift, a 23andMe Personal ***Genetic Genetic*** Information. To use our ***genetic*** testing services, you must purchase, or receive as a gift, a 23andMe Personal ***Genetic***

Sort by: Custom Ranking ▾

Crawl Dates

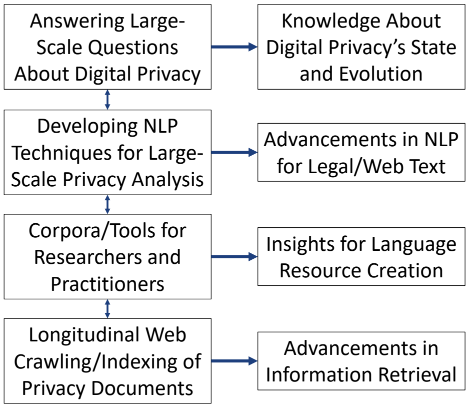
- ☐ July, 2019
- ☐ March, 2020
- ☐ March, 2021
- ☐ October, 2022

Filter

Flesch-Kincaid Reading Level (0 - 18): 0 | 18

Vagueness (0 - 1):

PrivaSeer Search interface


Answering Large-Scale Questions About Digital Privacy → Knowledge About Digital Privacy's State and Evolution

Developing NLP Techniques for Large-Scale Privacy Analysis → Advancements in NLP for Legal/Web Text

Corpora/Tools for Researchers and Practitioners → Insights for Language Resource Creation

Longitudinal Web Crawling/Indexing of Privacy Documents → Advancements in Information Retrieval

Overall project structure

Motivation and Goal

The wealth of privacy policies available on the web contrasts with the challenges of understanding the state of digital privacy at scale.

The PrivaSeer Project builds large-scale, longitudinal, annotated, and usable resources for the study of website privacy policies.

Resources

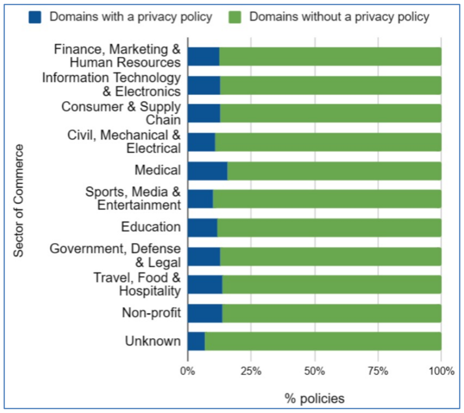
The **PrivaSeer Corpus** consists of 3,967,487 website privacy policies, complete with metadata such as sector of activity, readability metrics, and mentions to tracking technologies and regulations.

PrivaSeer Search is a search engine for exploring the corpus. This allows potential corpus users to test-drive the corpus prior to downloading it. It also makes the corpus accessible to privacy researchers with limited data analysis or programming skills.

The corpus and search engine are available to the public on the project website. The corpus is subject to a Creative Commons BY-NC-SA license.

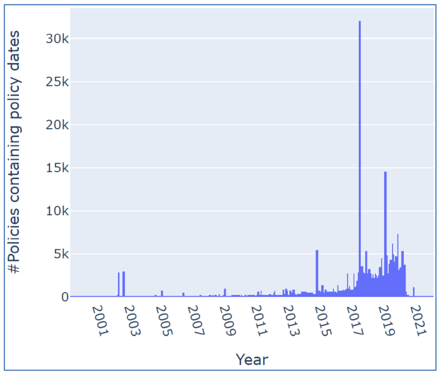
Analysis

See the project website for our publications in ACL, DocEng, ICWE, and PETS.



Percentages of domains with privacy policies by sector of commerce

Distribution of dates that appear in privacy policies (up to September 2021, as limited by data collection)



PrivaSeer | HLT LAB The Human Language Technologies Lab at Penn State | PennState College of Information Sciences and Technology | SCHOOL OF INFORMATION UNIVERSITY OF MICHIGAN | FUTURE OF PRIVACY FORUM | NSF