PrivaSeer

A Large-Scale, Longitudinal Resource to Advance Technical and Legal Understanding of Textual Privacy Information

Shomir Wilson & Lee Giles, Penn State
Florian Schaub, University of Michigan
Gabriela Zanfir-Fortuna, Future of Privacy Forum

PennState College of Information Sciences and Technology

UNIVERSITY OF MICHIGAN

FUTURE OF PRIVACY FORUM

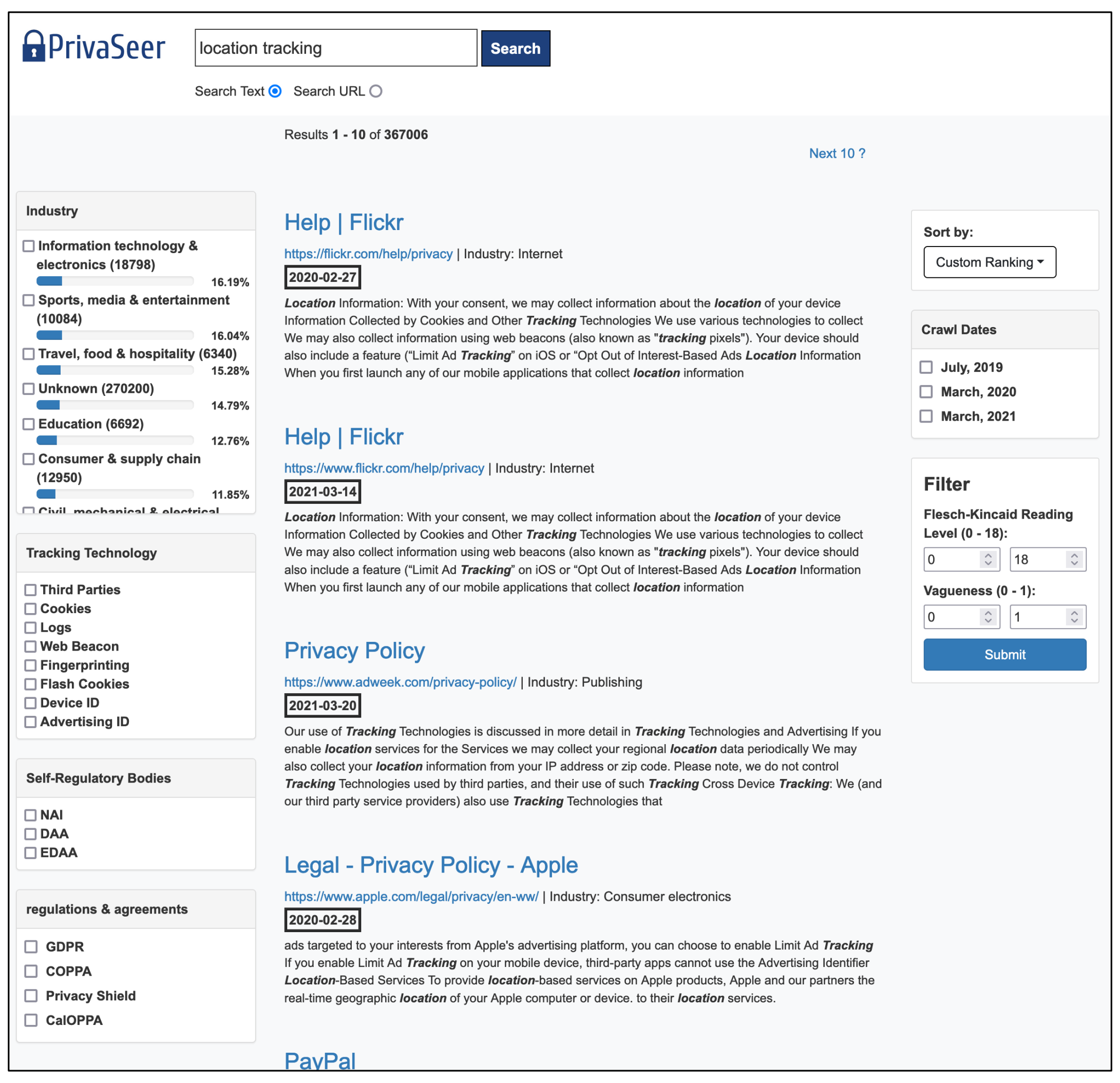
Try our search engine for over 1.4M privacy policies at privaseer.ist.psu.edu

We're building a **large-scale**, **annotated**, and **searchable resource** of **privacy-related texts**

- Advancing natural language processing techniques of privacy documents
- Facilitating research on privacy documents with infrastructure and tools
- Analyzing the state of privacy at scale

Challenges

- **A huge volume of privacy-related text exists online**: privacy policies, terms of use, cookie policies, do-not-sell notices, privacy laws, bills, regulatory guidance, etc.
- **Collecting and analyzing privacy text is necessary** for privacy research, practice, and effective regulation and enforcement.
- **Lack of infrastructure and tools** leads to repeated and duplicative effort across research projects.



PrivaSeer

location tracking

Search Text ● Search URL ○

Results **1 - 10** of **367006**

Industry

- ☐ Information technology & electronics (18798) 16.19%
- ☐ Sports, media & entertainment (10084) 16.04%
- ☐ Travel, food & hospitality (6340) 15.28%
- ☐ Unknown (270200) 14.79%
- ☐ Education (6692) 12.76%
- ☐ Consumer & supply chain (12950) 11.85%
- ☐ Civil, mechanical & electrical

Tracking Technology

- ☐ Third Parties
- ☐ Cookies
- ☐ Logs
- ☐ Web Beacon
- ☐ Fingerprinting
- ☐ Flash Cookies
- ☐ Device ID
- ☐ Advertising ID

Self-Regulatory Bodies

- ☐ NAI
- ☐ DAA
- ☐ EDAA

regulations & agreements

- ☐ GDPR
- ☐ COPPA
- ☐ Privacy Shield
- ☐ CalOPPA

Help | Flickr
https://flickr.com/help/privacy | Industry: Internet
2020-02-27
Location Information: With your consent, we may collect information about the ***location*** of your device Information Collected by Cookies and Other ***Tracking*** Technologies We use various technologies to collect We may also collect information using web beacons (also known as "***tracking*** pixels"). Your device should also include a feature ("Limit Ad ***Tracking***" on iOS or "Opt Out of Interest-Based Ads ***Location*** Information When you first launch any of our mobile applications that collect ***location*** information

Help | Flickr
https://www.flickr.com/help/privacy | Industry: Internet
2021-03-14
Location Information: With your consent, we may collect information about the ***location*** of your device Information Collected by Cookies and Other ***Tracking*** Technologies We use various technologies to collect We may also collect information using web beacons (also known as "***tracking*** pixels"). Your device should also include a feature ("Limit Ad ***Tracking***" on iOS or "Opt Out of Interest-Based Ads ***Location*** Information When you first launch any of our mobile applications that collect ***location*** information

Privacy Policy
https://www.adweek.com/privacy-policy/ | Industry: Publishing
2021-03-20
Our use of ***Tracking*** Technologies is discussed in more detail in ***Tracking*** Technologies and Advertising If you enable ***location*** services for the Services we may collect your regional ***location*** data periodically We may also collect your ***location*** information from your IP address or zip code. Please note, we do not control ***Tracking*** Technologies used by third parties, and their use of such ***Tracking*** Cross Device ***Tracking***: We (and our third party service providers) also use ***Tracking*** Technologies that

Legal - Privacy Policy - Apple
https://www.apple.com/legal/privacy/en-ww/ | Industry: Consumer electronics
2020-02-28
ads targeted to your interests from Apple's advertising platform, you can choose to enable Limit Ad ***Tracking*** If you enable Limit Ad ***Tracking*** on your mobile device, third-party apps cannot use the Advertising Identifier ***Location***-Based Services To provide ***location***-based services on Apple products, Apple and our partners the real-time geographic ***location*** of your Apple computer or device. to their ***location*** services.

PayPal

Sort by: Custom Ranking

Crawl Dates

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

Filter

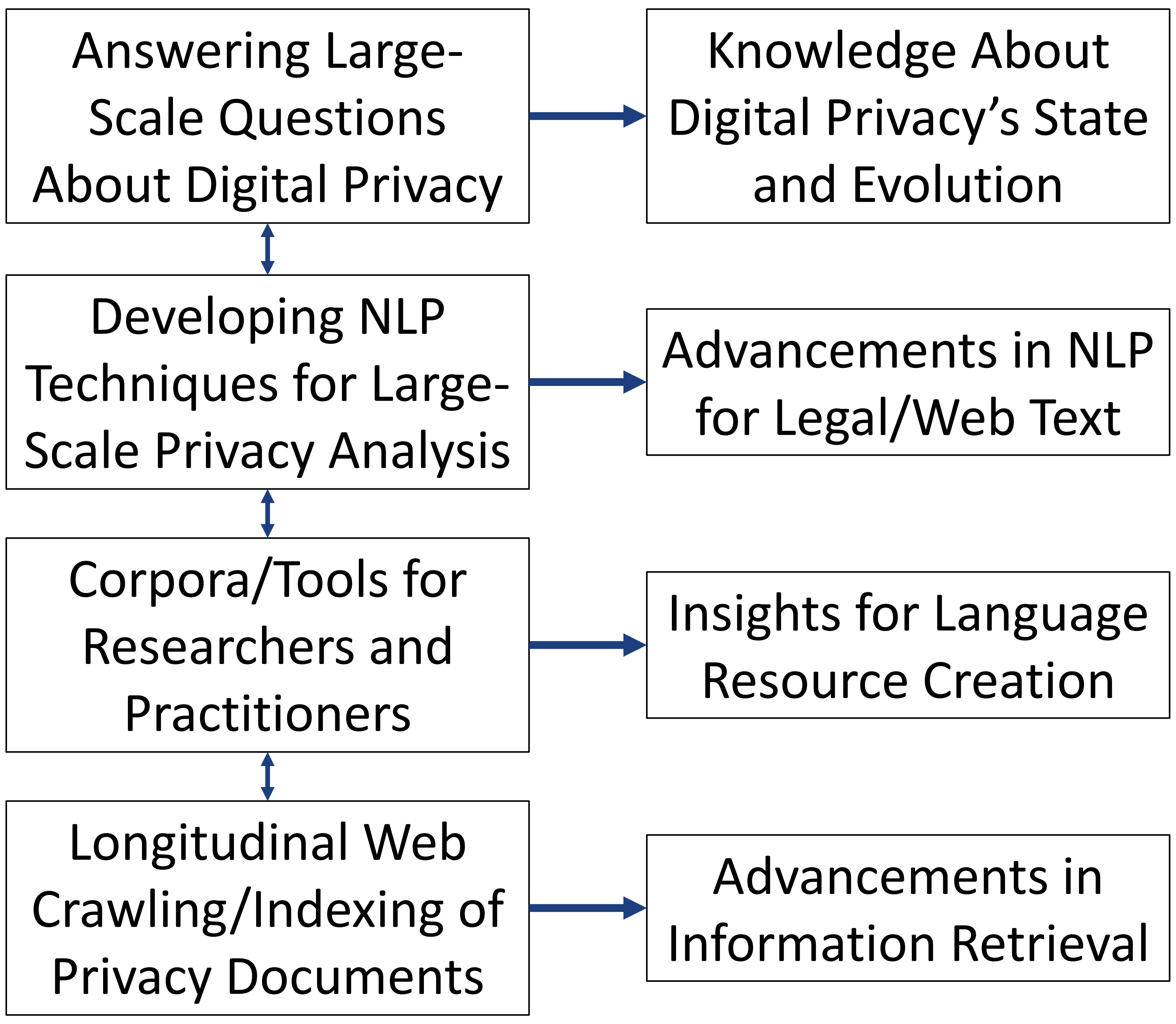
Flesch-Kincaid Reading Level (0 - 18): 0 – 18

Vagueness (0 - 1): 0 – 1

PrivaSeer search engine prototype (April 2022)

Project Objectives:

- **Enable analyses** of the state of privacy at an unprecedented scale
- **Remove barriers** for privacy analysis
- **Provide resources and tools** for researchers, practitioners, and policymakers



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

Broader Impacts

Research: Radically reduce effort needed to analyze privacy documents at scale.

Provide public tools and release corpora for research.

Industry & Public Policy: Enable data-driven and evidence-based decision and policy making.

Outreach to policymakers and practitioners.

Education & Broadening Participation: Train diverse team of students and fellows.

Foster diverse and interdisciplinary privacy research community.

This work is supported by NSF grants #2105736, #2105734, #2105745: Collaborative Research: SaTC: CORE: Medium: A Large-Scale, Longitudinal Resource to Advance Technical and Legal Understanding of Textual Privacy Information

The 5th NSF Secure and Trustworthy Cyberspace Principal Investigator Meeting (2022 SaTC PI Meeting)
June 1-2, 2022 I Arlington, Virginia