The Creation and Analysis of a Website Privacy Policy Corpus

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, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh

Motivation

Privacy policies are pervasive online, but they are **long and complex** documents that Internet users rarely read.

Internet users who are unaware of what happens to their data cannot make meaningful privacy choices.

We built an annotated corpus of privacy policies to enable NLP efforts to automate their interpretation. This was the first large-scale effort to annotate privacy policies at such a fine level of detail.

Corpus Composition

The annotations consist of **data practices** which fall into ten different **categories**. Each data practice is grounded in **spans** of policy text and has a category-specific set of **attributes**.

Documents	115
Words	266,713
Data Practices	23,194
Attr.-Value Selections	128,347
Text Span Selections	102,576
Annotators Per Policy	3
Annotators Total	10

Below: Data practice statistics for the entire corpus (frequency) and per policy (median). Fleiss' Kappa is calculated at the segment level.

Data Practice Category	Freq.	Median	Kappa
First Party Collection/Use	8,956	74	.76
Third Party Sharing/Collection	5,230	39	.76
Other	3,551	25	.49
User Choice/Control	1,791	13	.61
Data Security	1,009	7	.67
Int'l and Specific Audiences	941	6	.87
User Access, Edit, and Deletion	747	5	.74
Policy Change	550	4	.73
Data Retention	370	2	.55
Do Not Track	90	0	.91

Segment-Category Prediction

Predicting the practice categories of policy segments is a first step to simplifying or automating the annotation process.

	Log. Reg.			SVM		
Data Practice Category	P	R	F	P	R	F
First Party Collection/Use	.73	.67	.70	.76	.73	.75
Third Party Sharing/Collection	.64	.63	.63	.67	.73	.70
User Choice/Control	.45	.62	.52	.65	.58	.61
Across All 10 Categories	.53	.65	.58	.66	.66	.66

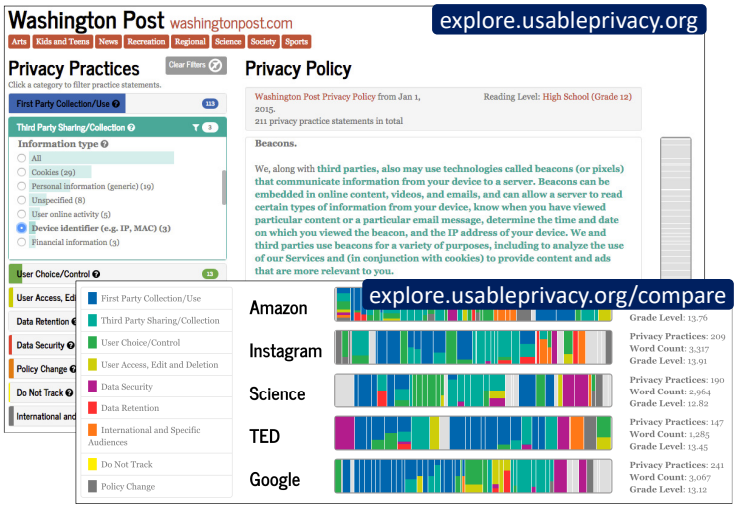
The Annotation Process

We worked with legal experts to develop an annotation scheme for privacy policy text. Skilled annotators used a **web-based annotation tool** (below) to apply it to policy text.

Policy text was divided into **segments** which were roughly equivalent to paragraphs. Annotators worked on one segment at a time.

A Resource for the Research Community

Researchers can explore the corpus on our interactive website.



You can download the corpus at data.usableprivacy.org

Future Directions

The corpus enables research in several directions, including:

- Automated extraction of data practices from text
- Cohesive interpretation of data practices in a privacy policy
- Identification of sectoral norms and outliers
- User-oriented summarization of privacy policies

USABLEPRIVACY.ORG the usable privacy policy project

NSF SaTC Frontier project
Lead PI: Norman Sadeh
sadeh@cs.cmu.edu

NSF · Carnegie Mellon University · Fordham University · Stanford University