# The Creation and Analysis of a Website Privacy Policy Corpus

Problem: There has been a relative dearth of datasets appropriate for identifying data practices in privacy policies.

This information consists of 23K data practices, 128K practice attributes, and 103K annotated text spans,

**Corpus Creation and Structure**

1. Privacy Policy Selection

Twostep process for policy selection:

* Relevance based website pre-selection
* sector-based sub-sampling

Selection Process:

* monitored Google Trends (Google, 2015) for one month (May 2015) to collect the top five search queries for each trend.

for each query we retrieved the first five websites listed on each of the first 10 pages of results.

Process (Simple version): One month Google Trends -> Each trend: Five search queries -> Each query: First five websites listed on the each of the first 10 pages of results

Result: a diverse sample of 1,799 unique websites.

* sub-sampled from this website dataset according to DMOZ.org’s top-level website sectors.2

organized the dataset into 15 sectors

excluded the “World” sector and limited the “Regional” sector to the “U.S.” subsector in order to ensure that all privacy policies in our corpus are subject to the same legal and regulatory requirements.

ranked the websites in each sector according to their frequency in the retrieved search results.

selected eight websites from each sector by randomly chosing two websites from each rank quartile.

Process (Simple version): Based on step 1 results -> sub-sampled websites into 15 sectors -> limited the “Regional” sector to the “US” -> Each sector: Rank websites -> Each rank quartile: Randomly selected two websites -> Each rank: Eight websites

Note:

* + Each website belongs to an American company, if not, randomly reselects in the same rank quartile.
  + Some privacy policies covered more than one of the selected websites: One company has several websites.

Result: 115 privacy policies across 15 sectors

1. Annotation Scheme and Process

a small group of domain experts (privacy experts, public policy experts, and legal scholars) identified different data practice categories and their descriptive attributes from multiple privacy policies.

Categories:

The final annotation scheme consists of ten data practice categories:

* First Party Collection/Use: how and why a service provider collects user information.
* Third Party Sharing/Collection: how user information may be shared with or collected by third parties.
* User Choice/Control: choices and control options available to users.
* User Access, Edit, & Deletion: if and how users may access, edit, or delete their information.
* Data Retention: how long user information is stored.
* Data Security: how user information is protected.
* Policy Change: if and how users will be informed about changes to the privacy policy.
* Do Not Track: if and how Do Not Track *signals3* for online tracking and advertising are honored.
* International & Specific Audiences: practices that pertain only to a specific group of users (e.g., children, Europeans, or California residents).
* Other: additional sub-labels for introductory or general text, contact information, and practices not covered by the other categories.

Process:

Definition:

An individual data practice

* belongs to one of the ten categories above
* articulated by a category-specific set of attributes
* e.g. a User Choice/Control data practice
  + four mandatory attributes (Choice Type, Choice Scope, Personal Information Type, Purpose)
  + one optional attribute (User Type)

The annotation scheme defines a set of potential values for each attribute

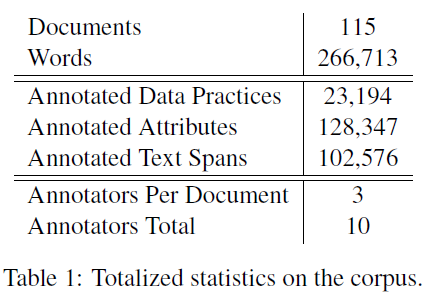
each attribute also may be associated with a text span in the privacy policy

The set of mandatory and optional attributes reflects the potential level of specificity with which a data practice of a given category may be described.

Unspecified, allowing annotators to express an absence of information

Annotation Process:

* Tool:



* Preparation:

privacy policies were divided into paragraph-length segments

annotators read in the tool, one at a time in sequence.

* Each segment:

An annotator may label 0 or more data practices from each category.

* Create a data practice:

Select a practice category -> Specify values and text spans for each of category’s attributes.

1. Composition of the OPP-115 Corpus (Online Privacy Policies, set of 115)

Policy Contents

Data practices are associated with 128K values for attributes and 103K selected spans of policy text.

Selection of a text span for mandatory attributes, did not require a text-based justification for optional attributes or attributes marked as “Unspecified”

Table 2 result:

First Party Collection/Use and Third Party Sharing/Collection dominated the rankings by frequency: the collection, usage, and sharing of user data are the primary concerns that compel the production of privacy policies.

Data practices in the Other category, while frequent, were mostly statements that were ostensibly not about user data

57% were introductory, contact, or generic information.

a practice category covers a policy segment if two of three annotators each identified at least one practice from that category in the segment text.

Differences in the category rankings by frequency and by coverage reveal that practices in some categories are less tightly clustered than others.

Data Retention is the second rarest practice category but ranks fourth by segment coverage.

Do Not Track and International and Specific Audiences remain standout categories with the greatest segment-level agreement.

Consolidating Annotators’ Work