Table: sites

Name	Type	Description
id	INTEGER PRIMARY KEY	ID of the website
domain	TEXT	Website's domain as extracted from the Alexa lists. May include subdomains.
categories	TEXT	Semicolon separated categories, as extracted from Webshrinker's API

Table: alexa_ranks

Name	Type	Description
site_id	INTEGER, FOREIGN KEY>sites (id)	ID of the website
year	TEXT	Year of capture
phase	TEXT	Phase of capture (A is Jan-June, B is July-Dec)
rank	INTEGER	Alexa rank at the time of the capture, may not be defined

Table: policy_texts

Name	Type	Description
id	INTEGER PRIMARY KEY	
policy_text	TEXT	Markdown formatted privacy policy text
flesch_kincaid	REAL	Flesch-Kincaid readability score
smog	REAL	SMOG readability score, based on a sample of 30 sentences
flesch_ease	TEXT	Flesch Ease readability score
length	INTEGER	Number of characters
sha1	TEXT	SHA1 hash of the text

simhash	TEXT	Simhash of the text

Table: policy_htmls

Name	Type	Description
id	INTEGER PRIMARY KEY	
policy_html	TEXT	The HTML of the whole privacy policy page
policy_html_sha1	TEXT	SHA1 of the policy HTML

Table: policy_reader_view_htmls

Name	Type	Description
id	INTEGER PRIMARY KEY	
policy_reader_view_html	TEXT	The HTML of the privacy policy only (without the boilerplate)
policy_reader_view_html_sha1	TEXT	SHA1 of the policy-only HTML

Table: policy_snapshots

Name	Type	Description
id	INTEGER PRIMARY KEY	
site_id	INTEGER	FOREIGN KEY sites (id)
homepage_snapshot_url	TEXT	The Internet Archive (IA) URL of the homepage
policy_snapshot_url	TEXT	The IA URL of the privacy policy
policy_url	TEXT	The URL of the privacy policy, without the IA prefix

homepage_snapshot_redirected_url	TEXT	The IA URL the homepage snapshot redirected to
year	INTEGER	Year of capture
phase	TEXT	Phase of capture (A/B)
policy_text_id	INTEGER	ID of the corresponding privacy policy text FOREIGN KEY policy_texts (id)
policy_html_id	INTEGER	ID of the corresponding privacy policy HTML FOREIGN KEY policy_htmls (id)
policy_reader_view_html_id	INTEGER	ID of the corresponding privacy policy-only HTML (without the boilerplate) FOREIGN KEY policy_reader_view_htmls (id)
file_type	TEXT	Filetype of the privacy policy: PDF or HTML
policy_title	TEXT	The title of the privacy policy
link_text	TEXT	The text of the link to the policy on the homepage
pdf_filename	TEXT	The filename of the PDF
classifier_probability	REAL	The probability this document is a privacy policy, as determined by our classifier
analysis_subcorpus	INTEGER	1 if the privacy policy snapshot is in the analysis subcorpus
parked_domain	INTEGER	1 if the privacy policy snapshot belongs to a parked domain
cross_domain_homepage_redir	INTEGER	1 if the homepage redirected to a page on another domain/origin