Free Law Project

http://freelawproject.org

Brian W. Carver

&

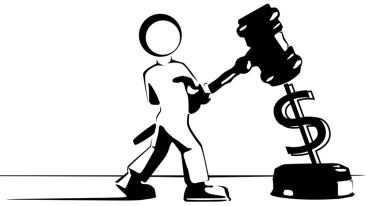
Michael Lissner

Columbia University Web Archiving Conference
June 4, 2015

@brianwc

@freelawproject

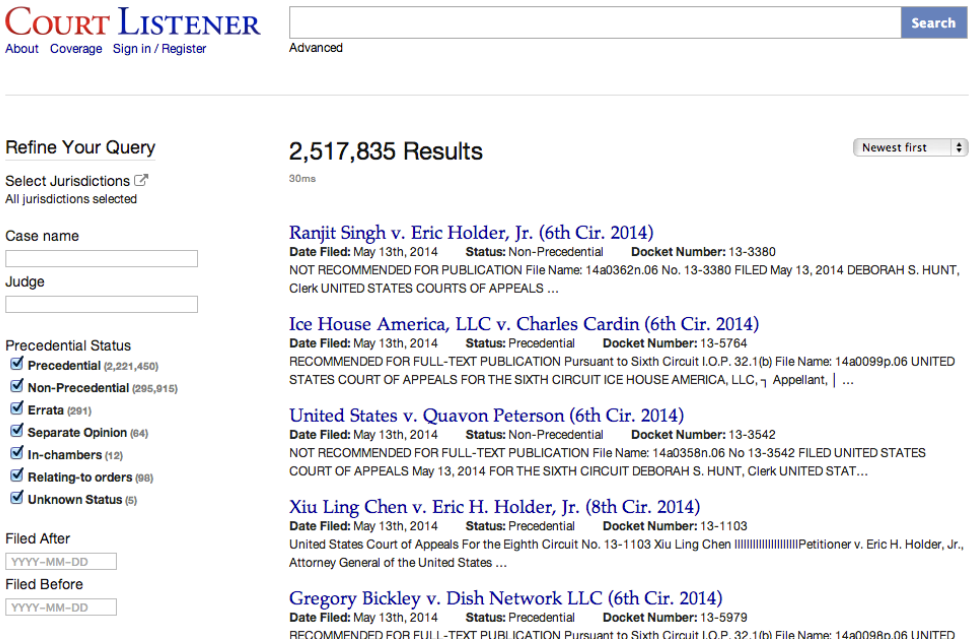
Free Law Project



- Free Law Project is a California non-profit that seeks to:
  - provide free online access to primary legal materials,
  - develop online legal research tools, and
  - support academic research on legal corpora.
  -
- Co-founders:
  - Michael Lissner &
  - Brian W. Carver

Free Law Project

- Free Law Project sponsors the development of:
  - CourtListener, https://www.courtlistener.com



Free Law Project

- Free Law Project sponsors the development of:
  - Juriscraper, https://github.com/freelawproject/juriscraper, a software system for harvesting the court opinions provided on appellate court websites.
    - Covers all federal circuit courts and Supreme Court.
    - Now covers all state courts of last resort and most all intermediate appellate state courts.
    - Covers various special jurisdictions, such as Court of Federal Claims, Veterans Claims, Tax, Int'l Trade, etc.
  - This amounts to watching over 200 different webpages for new documents, every weekday.

Free Law Project

•Free Law Project partners with Princeton's Center for Information Technology Policy on:

•RECAP, http://recapthelaw.org, browser extensions that improve the experience of browsing the federal electronic document system, PACER, while helping PACER users build a free and open repository of public court records.

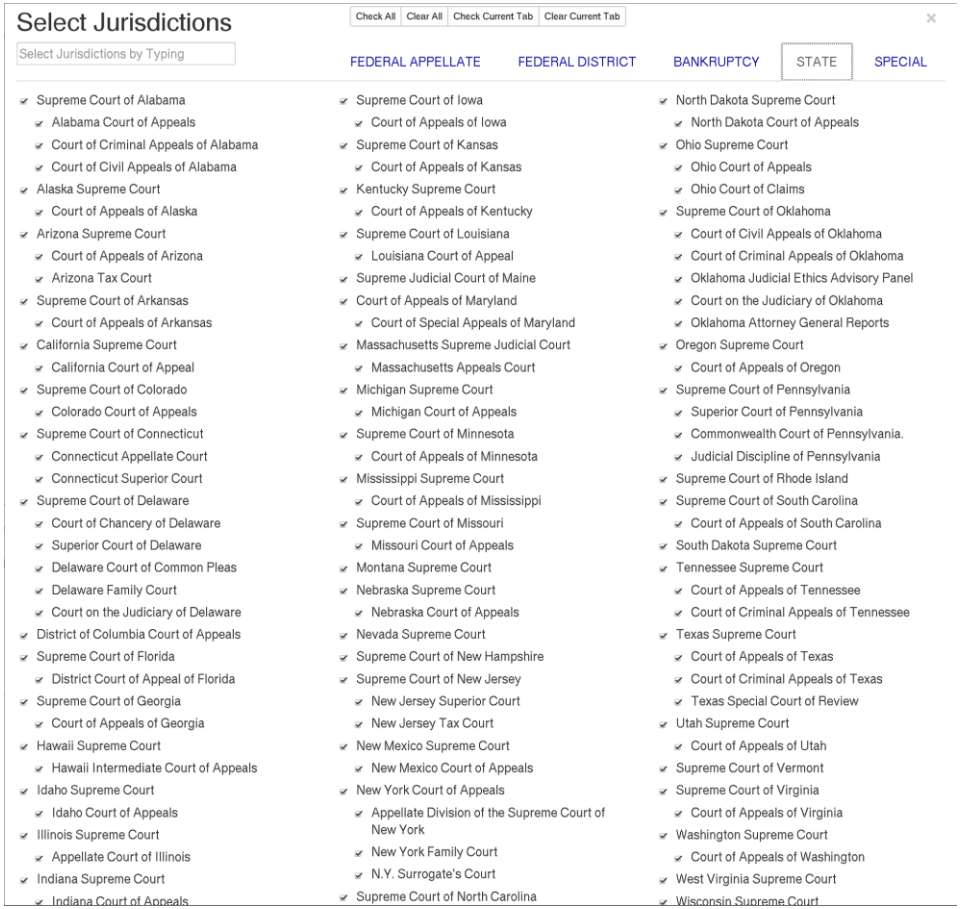
•Michael will give a lightning talk about RECAP tomorrow.

Free Law Project

- Free Law Project sponsors the development of:
  - Online legal research tools on the CourtListener platform, such as a citator, table of authorities for each document, sophisticated search filtering options, improvements to relevancy search results, etc.
  - An Application Programming Interface (API) to provide access to all court opinions and key related metadata in our database, to enable others to programmatically utilize our collections.

Free Law Project

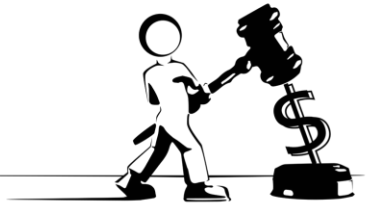
- What did you do with our money?!



Free Law Project

- Scrapers
  - We want the metadata about the documents we collect to be accurate.
  - A one-size-fits-all solution cannot guarantee this across hundreds of different court websites.
  - We do have an extensible general core.
  - The details for each site are handled independently, and this creates a very low barrier to entry for new developer-contributors.

Free Law Project



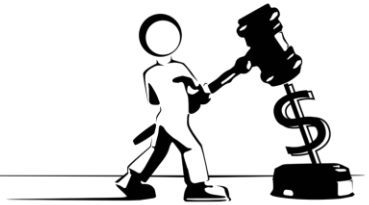
56 lines (45 sloc) | 1.717 kB — Raw | Blame | History

```
"""Scraper for Supreme Court of Maine
CourtID: me
Court Short Name: Me.
Author: Brian W. Carver
Date created: June 20, 2014

History:
  2014-06-25 (est): Added code for additional date formats.
  2014-07-02: Was receiving InsanityException and tweaked date code to get some
              missing dates.
  2014-12-15: Fixes insanity exception by tweaking the XPaths.
"""

from datetime import datetime

from lxml import html
from juriscraper.OpinionSite import OpinionSite


class Site(OpinionSite):
    def __init__(self):
        super(Site, self).__init__()
        self.court_id = self.__module__
        self.url = 'http://www.courts.maine.gov/opinions_orders/supreme/publishedopinions.shtml'

    def _get_download_urls(self):
        path = '//table//tr/td[2]/a[1]/@href'
        return list(self.html.xpath(path))

    def _get_case_names(self):
        case_names = []
        for e in self.html.xpath('//table//tr/td[2]/a[1]'):
            s = html.tostring(e, method='text', encoding='unicode')
            case_names.append(s)
        return case_names

    def _get_case_dates(self):
        path = '//table//tr/td[3]/text()'
        date_styles = ['%B %d, %Y', '%B %d,%Y']
        dates = []
        for s in self.html.xpath(path):
            for date_style in date_styles:
                try:
                    d = datetime.strptime(s.strip(), date_style).date()
                except ValueError:
                    continue
                dates.append(d)
        return dates
```

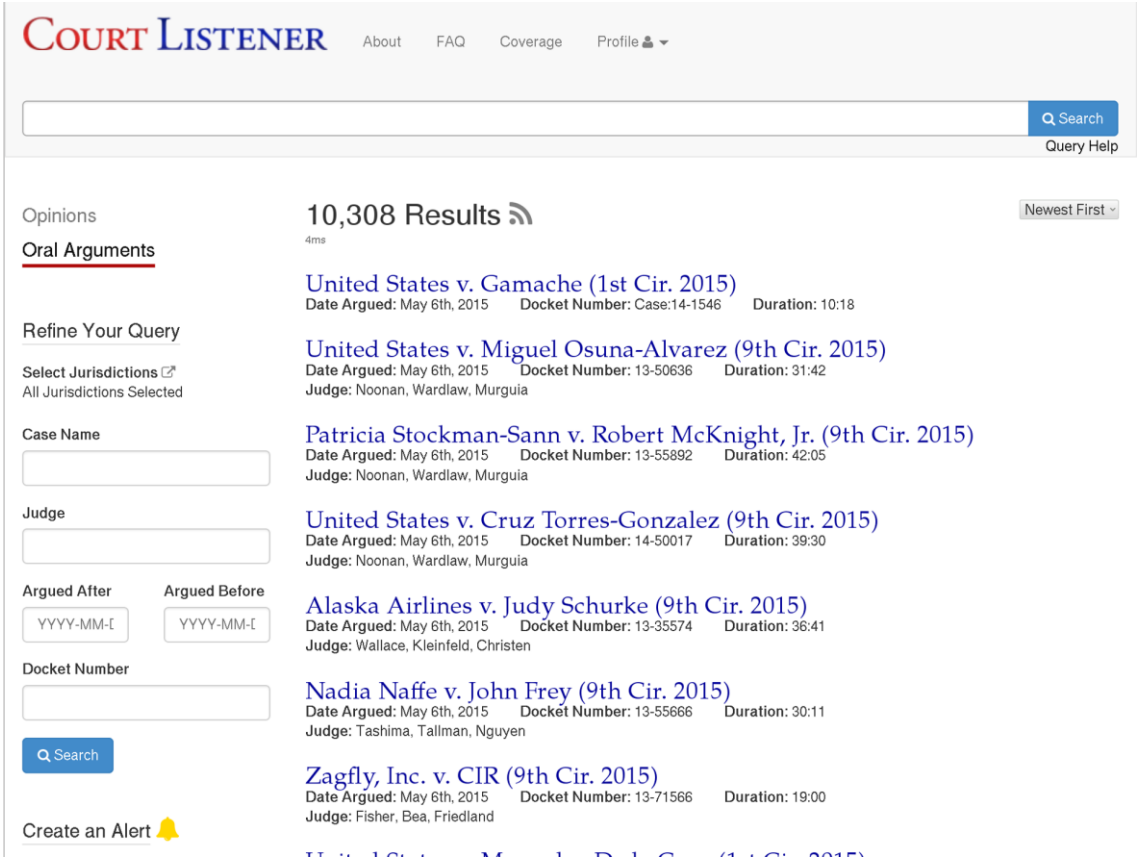
Free Law Project

- Scrapers
  - When we started this grant, we had 87 state court scrapers.
  - When we finished work under this grant, we had 154 state court scrapers.
  - Since completing this work we have added over 58,000 state court opinions to our archive.

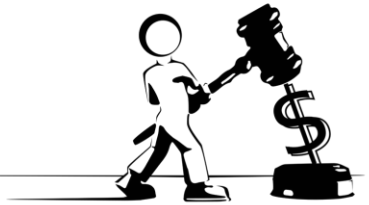
Free Law Project



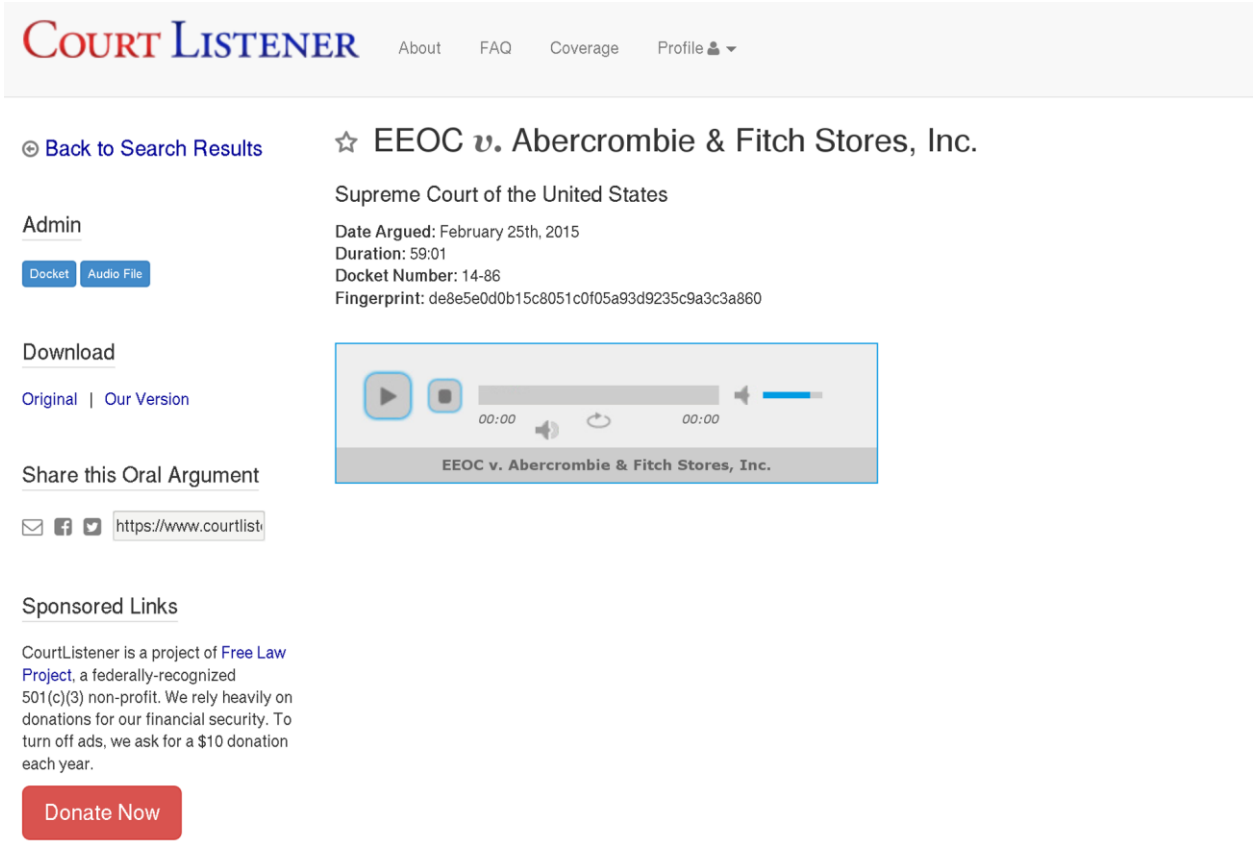
- What did you do with our money?!



Free Law Project



- What did you do with our money?!



Free Law Project

- Oral Arguments
  - We preserve the original audio file.
  - We create an improved .mp3 with metadata, album covers, etc.
  - These are also accessible via our API and bulk downloads.

Free Law Project

- Oral Arguments
  - When we started this grant, we had 0 oral argument audio files.
  - Today we host more than 10,000 oral argument audio files. (257 *days* of listening.)
  - By adding a new document/entity type we began the process of generalizing our infrastructure to accommodate new document types in the future.
  - DC Legal Hackers named it a Legal Hack of the Year for 2014.

Free Law Project

- What did we do under this grant?
  - Complete our daily coverage of all 50 states, even the state courts with terrible websites, by building lots of web scrapers.
  - Add oral argument audio scrapers for the federal appellate courts.
  - Add an open source audio player to our site.
  - Design our system to recognize a new type of document/media.

Free Law Project

- Pain Points
  - PDFs are horrible.
  - Would like to automate a no-cost audio-to-text transcription solution. Does it exist?
  - Ongoing maintenance and monitoring.
- Future (Related) Work
  - Increasingly, courts are providing oral argument *video*.

Free Law Project

- Thinking beyond the law:
  - Juriscraper has taught us a lot about how to harvest material from lots of websites.
  - One could change some variable names and use the same library to harvest anything.
  - Harvesting audio taught us a lot about audio processing and audio metadata that is useful to anyone archiving audio files.

Free Law Project

- Get in touch!
  - http://freelawproject.org
  - https://www.courtlistener.com
  - http://recapthelaw.org
  - @freelawproject
  - @courtlistener
  - @recapthelaw
  - @brianwc
  - @mlissner