UC Berkeley School of Information (South Hall)
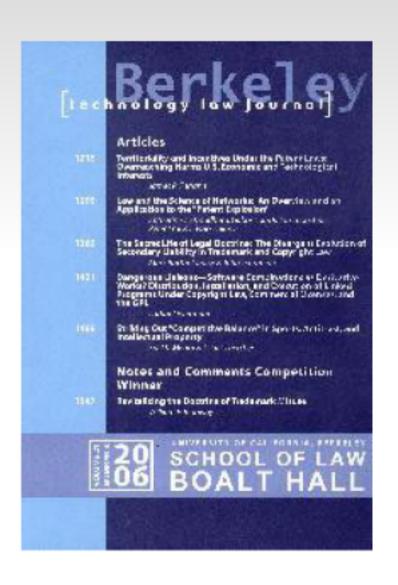
- *bcarver@ischool.berkeley.edu*
- @brianwc

Brian W. Carver

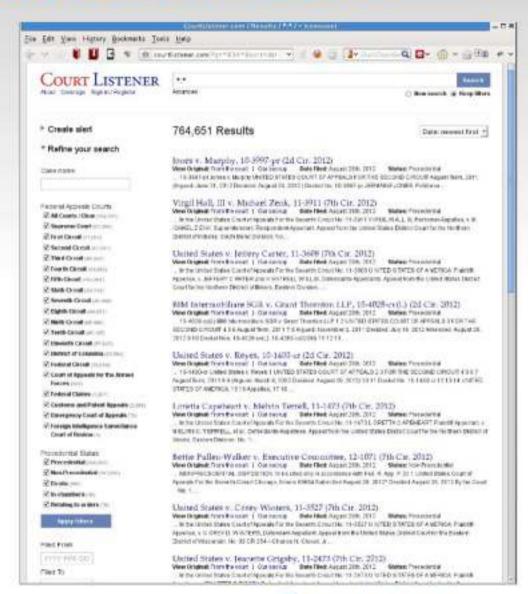UC Berkeley School of **Information**

My Research & Teaching

- Copyright & Cyberlaw

- Free software and other forms of commons-based peer production

- Technology and innovation policy

- Open educational resources

- Public access to the law

Brian W. Carver

CourtListener.com

- *Michael Lissner*

Brian W. Carver

CourtListener.com

- Motivations:
  - Daily Awareness Tool (researchers, journalists, lawyers)
  - Public Access to the Law
    - The *complete* case law corpus is simply not online without enormous expense (yet!)
  - Survivability/Persistence (The "If I get hit by a bus" test)
    - All the code is under open source licenses on bitbucket.org
    - All the documents can be downloaded by anyone
  - General purpose case law research tool and citator
  - Give it all away for free

Features of our corpus

- 765,000+ documents (with more added daily)

- Coverage of the federal appellate courts, *i.e.*, the 13 circuit courts within the federal system and the Supreme Court of the United States (with more jurisdictions, e.g., the 50 states, to be added soon).

- The entire Supreme Court corpus from 1754 to the present. (From 1 U.S. 1 on; 61,254 documents)

- A (mostly complete) Circuit Court corpus from the late 1940s to present.

- Inter-linked citations.

Features of our corpus

- Court opinions are in the public domain, *i.e.*, we and researchers using the corpus, do not have to think about copyright issues!

- *You* can download all our documents at no cost! (Please don't everyone try this at once. 3.3GB+)

- Key metadata marked up in XML:

  - Court name, filing date, case name, our source, OCR?, document_id, doc_ids of citing cases, etc.

Download our corpus!

- For example, let's look at ways to access the First Circuit of Appeals (ca1):

    - All ca1 cases from 2009:
      http://courtlistener.com/dump-api/2009/ca1.xml.gz

    - All ca1 cases from June 2009:
      http://courtlistener.com/dump-api/2009/06/ca1.xml.gz

    - All ca1 cases from June 9, 2009:
      http://courtlistener.com/dump-api/2009/06/09/ca1.xml.gz

    - All cases for a given time period:
      http://courtlistener.com/dump-api/2009/06/09/all.xml.gz
      returns all of the cases from June 9, 2009 (all courts).

What's this look like?

Collecting Legal Decisions

- In August 2012 we added 2,195 documents, or about 100 documents added each weekday.

- This is from the federal appellate courts. We are in the process of adding all the state appellate courts and the federal district courts and that will increase the document count significantly.

- Unique web scrapers for each court, often multiple scrapers for the same court!

- Unified in our "Juriscraper" project at bitbucket.org

Legal Citators

- What is a citator?

  - Lexis' Shepard's ("shepardizing a case")

  - Westlaw's KeyCite

  - Bloomberg Law's BCite

- Given a reference of a legal decision, a citator allows the researcher to find newer documents which cite the original document and thus to reconstruct the judicial history of cases and statutes.

Legal Citators

- A citator is important because it can tell you things like the history of the case on appeal, the treatment by subsequent courts of legal issues in the case, the case's current validity, *i.e.*, is the case still good law?, and allows you to expand your research by pointing to other cases and secondary sources that cite the case.

- The other comprehensive no-cost online repositories of case law do not provide a citator. Google Scholar has a "How cited" tab that provides some quotations but the results a) are not customizable b) provide no information about how complete they might be and c) provide only extracted quotations without characterizing the citation.

Legal Citators

- Identifying citations is easy, right?
  - Ingle v. Landis Tool Co. (C.C.A.) 272 F. 464
  - Ingle v. Landis Tool Co., 272 Fed. 464 (3d Cir. 1921)
  - Ingle v. Landis Tool Co., (CCA 3d Cir. 1921) 272 Fed. 464
  - Ingle v. Landis Tool Co., 272 F. 464 (3d Cir. 1921)
  - Ingle v. Landis Tool Co., 272 Fed. 464 (3rd Circ., 1921)
  - Ingle v. Landis Tool Co., 272 Fed. 464 (C. C. A. 3d, 1921)
  - Ingle v. Landis Tool Co., 272 _____ LINE BREAK _____ F. 464 (3d Cir. 1921)

Legal Citations

- Two problems:

  - Standardization of citation formats (The BlueBook)

  - Adopting a citation format that can be known upon the first publication of a **digital** document by the Court, rather than a format that can only be known once West Publishing decides to place it in a bound **print** volume.

    - Some jurisdictions have adopted "neutral" citation formats, but there are lots of jurisdictions...

Our Citator

- Rowyn McDonald and Karen Rustad

- Added 4.2 million citation links among the 750,000+ documents

- Automatically detects and links citations every day as new documents are added.

- The citation relationships are part of the XML data dumps.

Our Citator

- When the citator work was completed it occurred to me that I could ask (and answer) a question that previously would have required the permission of Westlaw or Lexis:

  - What is the *most* cited case?

- One MySQL query later, the answer:

  - *Strickland v. Washington*, 466 U.S. 668 (1984).

  - A case on the ineffective assistance of legal counsel in death penalty cases.

Distributing Legal Corpora

- How do you pay for all this?
  - We don't.
  - My volunteer time; my students' volunteer time
  - Small grant from Berkeley School of Information
  - Some funding from Carl Malamud of resource.org
  - Out of my own pocket
- So far, so good, but we'd get further faster with a larger source of longer-term funding.

Research on a Legal Corpus

- What can we do with this corpus?

  - *You* tell me!

- Yesterday's presentation on personality using word clouds.

  - Idea: Compare Justice Scalia's cloud to Justice Kagan's.

Research on a Legal Corpus

- First Circuit cases from June 9, 2009:

Research on a Legal Corpus

- What can we do with this corpus?

- Matthew Jockers does work on 18th and 19th century literature determining authorial style: Jane Austen is very consistent; Charles Dickens is not.

  - Idea: Use these techniques to compare a judge's style over time or estimate authorship of *per curiam* opinions.

Research on a Legal Corpus

- What can we do with this corpus?

- My colleague Geoff Nunberg likes to trace the use of specific words over time.

  - Idea: Our Supreme Court corpus presents over 250 years of American English usage (in a niche domain).

- Main point:

  - West and Lexis have held legal corpora privately for so long that there is a lot of low-hanging fruit in doing computational analyses of legal corpora.

Questions?

- *bcarver@ischool.berkeley.edu*

 @brianwc

Brian W. Carver