The Berkeley National Reporter: Building a Free, Open Source Legal Citator

Rowyn McDonald
Karen Rustad

Masters Final Project Report
School of Information
University of California, Berkeley
May 2012

Table of Contents

Introduction

For our final project, our group built a free, functional, open source citator: a tool for identifying and tracking legal citations in a corpus of United States court decision text.

In this paper, we describe our problem scope, domain research, design decisions, development process for the citation-finding algorithm, results and metrics for the citator, and future directions for the citator's development.

The Problem

The Tyranny of Print

The legal field is known for having a technology suite well behind the curve compared to other industries. The reasons for this are beyond the scope of this paper. One contributing factor, however, is that legal doctrine's core information organization schema is inherently tied to the print form factor. To identify a particular court decision in the body of US case law (a source at least as important as legislative statutes in terms of figuring out what the law of the land actually is), you have to know which of West Publishing's court reporters covers it, which volume the case is in, and the first page on which the case is printed. These citations are completely arbitrary; none of these pieces of information (other than the reporter, based on the jurisdiction of the case) can be predicted or extrapolated from the text of the case or metadata thereof. When the US Supreme Court first hands down a new decision, that decision does not have an official citation and there is no canonical way to refer to it. Until West puts that decision into a physical book and prints it, the case practically does not exist.

Few lawyers still have shelves of physical case reporter volumes lining their walls in which they look up cases. Nearly all lawyers today use either West's online service Westlaw or LEXIS to match citations to cases and otherwise conduct legal research. However, the dominant legal citation format still reflects the old system of physical, printed court reporters and remains arbitrarily assigned by West. As a result, although the courts have found that West citations themselves are not copyrightable works[1], the dominant case identification system in the United States is effectively the property of a private company. This legal information organization roadblock makes building new, innovative applications involving U.S. case law unappealingly complicated for most programmers and startups.

Information Justice for Judicial Information

The U.S. case law's archaic information organization system is a significant obstacle for improving citizens' access to the law that governs them. It is an axiom of any democratic society that citizens

[1] Bender *v*. West, 158 F.3d 674 (2nd Cir. 1998)

must be able to read and learn about the law, including both statutes and case law. Free or low-cost access to court decisions is important for students, journalists, small-firm lawyers, and the people they serve to understand how the law of the land is interpreted and how it affects them.

Court decisions are in the public domain by default as they are governmental works.[2] However, the government itself does not provide access to them. When courts publish new decisions on their websites, they are typically published as PDFs (sometimes scanned images rather than extractable text) and older decisions are often cycled out and taken down. The courts themselves do not know what the unique, canonical identifier or citation for the decisions they hand down will be; decisions have only a docket number which is often inconsistently formatted and may be used to refer to multiple cases. We agree with Ian Gallacher that the reporter-dependent bibliographic system "has the unintended consequence of restricting free and open access to the law."[3]

Right now, the only way to access a complete repository of U.S. case law including canonical citations is through two legal research services: Westlaw and LexisNexis (which licenses West's citation schema). These services are extremely expensive. Westlaw engages in heavy price discrimination and makes it nearly impossible to compare prices between different plans or clients.[4] For a tiny, two-person law firm, an all-you-can-eat Westlaw Next plan costs $1000 per month. A document from 2009 about the classic Westlaw service lists per-minute costs from $9 to $24 and per-search query costs of $61 to $194, depending on the jurisdiction database.[5] The hourly rate for case law on Westlaw Next appears to be a flat $800 per hour; other databases, such as appellate court briefs, are north of $3000 per hour.[6] Unless you are at a law school or large law firm that can afford an all-you-can-eat plan, browsing US Supreme Court case law for a half hour sets you back about $400 on Westlaw Next and $720 on Westlaw--just to read documents that are in the public domain! At these prices, these services cannot be considered to provide public access to the law by any stretch of the imagination.

Legal Research Tools for All

Fortunately, most U.S. case law is now publicly available in some form. Public.Resource.Org serves large archive files of older federal court decisions; a few free-to-the-public services such as Google Scholar and Justia present this corpus of case law in a reasonably readable form. CourtListener.com,

---

[2] 17 U.S.C. §105; *see also* Carl Malamud, Three Revolutions in American Law (2009)

[3] Gallacher, I. "Cite Unseen: How Neutral Citation and America's Law Schools Can Cure Our Strange Devotion to Bibliographical Orthodoxy and the Constriction of Open and Equal Access to the Law." 70 Albany Law Review 491 (2007), 510.

[4] We found it next to impossible to check the pricing for multiple all-you-can-eat scenarios because the Westlaw store portal remembers the first practice type and employee count that you enter into the system and refuses to let you change it.

[5] http://library.law.emory.edu/fileadmin/library/Internet_Legal_Research_Guide/Cost_of_Westlaw_2009.ppt

[6] Lambert, Greg. "WestlawNext Pricing - Up To $3400 Per Hour!!" http://www.geeklawblog.com/2010/03/westlawnext-pricing-up-to-3400-per-hour.html 15 Mar. 2010.

originally developed by I School alumnus Michael Lissner as a free platform for staying abreast of recent court opinions, crawls court websites to add copies of newer decisions to public.resource.org's corpus. However, none of these free resources offers a *citator*—a tool that identifies:

- what other cases cite a given case C;
- to what *extent* the citing cases cite C (offhand citations versus extended discussion and analysis, often called "depth of treatment");
- whether that discussion of C is positive (upholding C's doctrine) or negative (overturning it), a process called "shepardizing" in legal jargon; and
- through the above, suggest whether C is still "good law" or not.[7]

Citators are fundamental to legal research because U.S. jurisprudence is founded on the principle of *stare decisis*: the idea that court cases are decided based on doctrines and decisions a court has made in the past. It is important when making a well-grounded legal argument that one cite past cases that serve as precedent for the argument. If your argument is based on a case that has since been overruled or overturned, your argument will almost certainly fail. Citators are also useful for finding cases that cover related topics or are relevant to a particular area of research. Finally—as in analysis of academic citation networks—a map of citations could help legal scholars identify influential decisions, judges, and courts over time and track the flow of information and ideas through the legal system.

Without a free, publicly accessible citator, public portals such as CourtListener are incomplete — people can read court decisions there but do not get an understanding of their context. To fix this, our group set out to build a free, open source citator tool on top of CourtListener's code base and case law corpus.[8]

[7] Google Scholar recently added some citator functionality to its legal search. However, since Google Scholar neither is open source nor includes its case law search in its API, there is no way for the public to build upon this service or count on its continued existence or development.

[8] The CourtListener code base, including our modifications, is publicly available at https://bitbucket.org/mlissner/search-and-awareness-platform-courtlistener/.

The Research

Conceptions of Case Law Information Organization Systems

A large share of the legal information access literature focuses on the problem of changing the case law citation standard to be independent of a single vendor and untethered from physical media. A rival, neutral standard, called "medium-neutral" citation, has been endorsed by the American Bar Association and adopted by the North Dakota Supreme Court and a few other state courts.[9] While a neutral, platform-agnostic citation format would be a clear improvement on the current model from a pure information organization perspective, and also reinforce the notion that the body of case law belongs to the public, none of the major federal courts has adopted the standard.[10] Adoption in the legal community is hindered both by institutional inertia and by the lack of support for medium-neutral citations in major legal research tools such as Westlaw (Gallacher 528). Furthermore, even if every court adopted a neutral citation format henceforth, backwards compatibility and the ability to parse older cases' citations would still a critical problem. Thus, we did not deem medium-neutral citation support to be a priority for our citator.

We also found papers on legal entity extraction: using computers to automatically determine the topic of a given legal document. Suggesting topics or categories of law that a case discusses and allowing users to filter by topic would be a useful feature for a citator to have. It was also interesting to learn that—for the discovery process, at least—a rules-based algorithm performed better than a Bayesian approach to identifying topics; this suggests that a Bayesian approach might not work as well for citation identification or analysis either.[11] We excluded topic extraction from the scope of our project.

In terms of a more general model for case law information, Bommarito et al conceived of the case law citation network as an "acyclic digraph"—that is, all the citations are directional and unlooping and the time dimension is crucial to the graph's structure.[12] A case can only cite cases older than it; it is impossible to cite cases that have not been decided yet. The paper's conceptual strategy is to find "sinks" (novel legal ideas) as they appear in early U.S. Supreme Court cases and observe how these ideas are recombined over time.

While we were not able to repurpose any of Bommarito's citation-finding work (the only deliverables in the paper were some information visualization graphics; the code that made them was not published), we did find the idea of case law as a series of incremental deltas on old ideas and doctrines compelling. It suggested that evaluating whether a given court decision is "good law"

---

[9] Martin, Peter W. Introduction to Basic Legal Citation (online ed. 2011). http://www.law.cornell.edu/citation/1-500.htm

[10] Minick, Courtney. "Public Domain Legal Citations." http://onward.justia.com/2010/12/17/public-domain-legal-citations/ 17 Dec 2010.

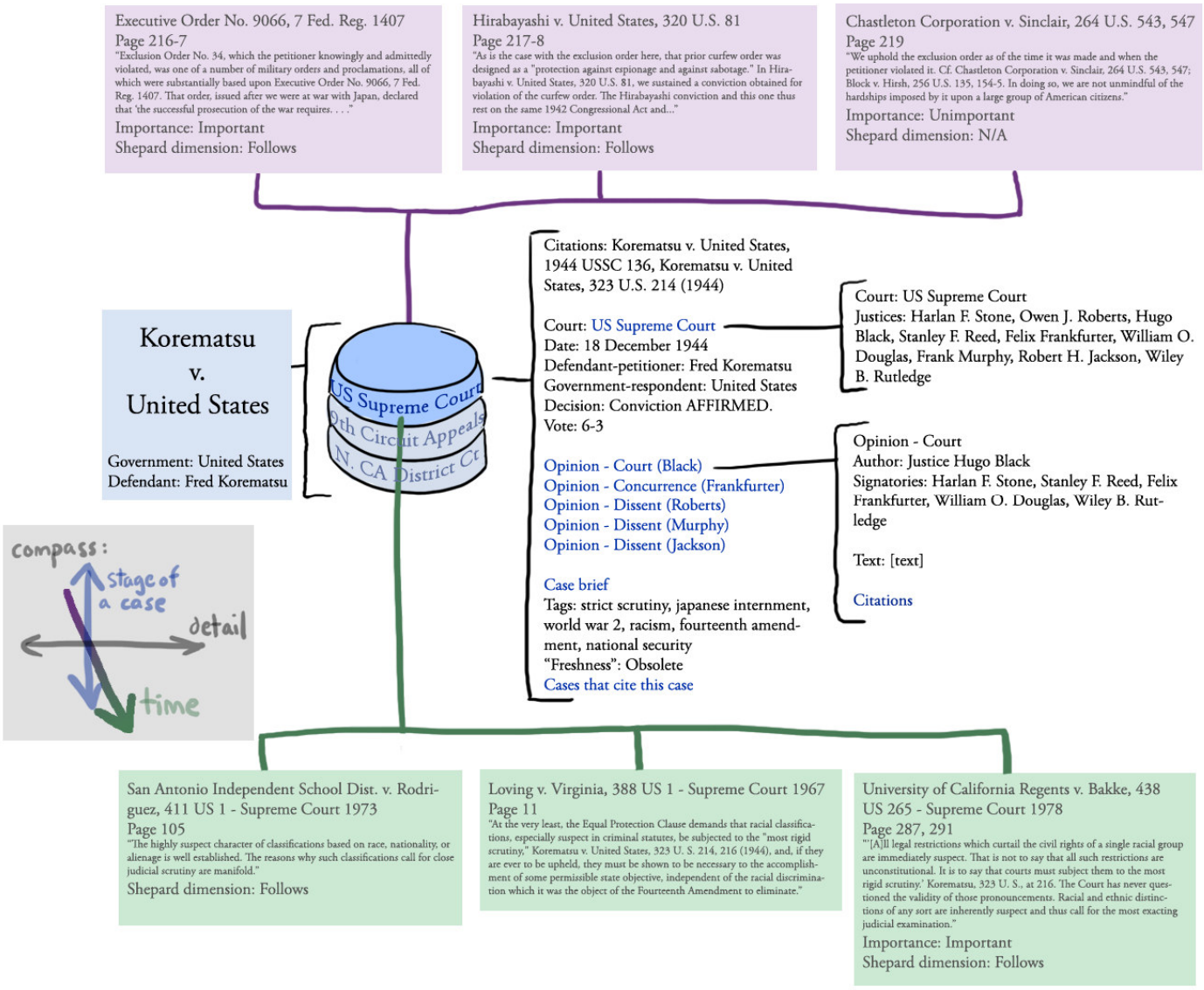
[11] Deshpande, Prasad et al. "Automated Concept Extraction to aid Legal eDiscovery Review". COMAD 2009, Mysore, India.

[12] Bommarito, et al. "Distance Measure for Dynamic Citation Networks." arXiv:0909.1819v3 [physics.soc-ph]

or not is insufficient; rather, you would prefer to evaluate sections, paragraphs, or individual lines for their grounding. Part of a case, or even the case's conclusion, might be invalidated by a later court decision, but other ideas from the case might still live on and be "good law". An example of this is the World War II case Korematsu *v*. United States. The majority opinion in the case found that Japanese internment was lawful because national security concerns during the war were sufficient to justify mass, race-based violations of the Fourth and Fifth Amendments. This finding has since been overturned and is clearly bad law. However, one piece of *Korematsu* still lives on. *Korematsu* was the first U.S. Supreme Court case to establish that laws establishing discrimination or unequal treatment under the law based on race or ethnicity were subject to "strict scrutiny"—the highest standard of inquiry and skepticism that the U.S. Supreme Court may apply. This precedent was later used in Brown *v.* Board, Bakke *v.* UC Regents, and other civil rights cases to justify the Supreme Court's power to nullify racially discriminatory laws or policies. In this project, we did not have time to model this more complicated view of case "freshness" in software, but we did envision how it might be reflected in the case reader interface.

A previous paper by co-author Karen Rustad proposed a comprehensive data model for organizing and browsing between cases.[13] In the "DRAPESH" ("Shepard" backwards phonetically) model, the unit of analysis is a Case, which consists of a stack of Trial Nodes each representing a court decision at a particular level in the judicial hierarchy. Trial Nodes have Citation relations to parent and child Cases and a variety of metadata, including links to relevant Courts and Judges and a set of Opinions.

[13] Rustad, Karen. "DRAPESH: Toward a Free, Multidimensional, Shepardizing Case Law Data Model." May 2011.



DRAPESH data model illustration, using Korematsu v. United States as an example. Figure from Rustad (2011), Appendix A.

The DRAPESH proposal went above and beyond the needs of a basic citator. Such a data model would not only create the citing and cited-by relations between cases that a citator relies on, but would also make it easy to answer questions about the decisions and citation behavior of particular courts and judges and about the judicial history and path through the court system of a given case. It also anticipated including citations to statutes and Congressional reports, not just cases. Thus, while DRAPESH was influential in the design of our citator (and ideas for visualizing citation data), implementing this model was also outside the scope for our project.

Service Design and Competing Services

During our research, we discovered several startups and projects that had tried and failed to maintain a free legal citator. Two examples:

- **PreCYdent:** PreCYdent was a free legal search tool that determined the importance, influence, and relevance of cases based on their citations in law review articles as well as in an incomplete corpus of case law. Apparently its proprietary algorithm got better results than Lexis or Westlaw at the time; positive reviews of the service abound on legal blogs.[14] Nonetheless, the service only lasted from 2006 to 2009; the founders had to shut it down due to lack of funds.[15]
- **AltLaw**: AltLaw was a collaboration between the law schools of Columbia and University of Colorado. When Google Scholar introduced case law search, AltLaw shut itself down as it believed its mission to make legal information publicly accessible was complete.[16] However, unlike Google Scholar or any other service, AltLaw had a public API that included citator information, which others could build upon. Despite the fact that Google Scholar generally has an API, four years of feature requests have failed to convince Google to extend the API to the legal search portion of the site.[17]

Surviving legal startups either charged for access to their database, did not provide a citator service, or pivoted away from competing in the general legal research market to serving a particular legal industry niche.[18]

Gallacher argues that "rather than another for-profit publisher seeking to make money from publishing the law, what is needed is a publisher with no profit motive, but that still has the resources to commit to such a project, and which has a scholarly interest in, and intellectual commitment to, the law for its own sake" (530). The number of for-profit startups that have attempted and failed to displace Westlaw and LexisNexis supports this. If we are to have a case law reporter service that is free, publicly-accessible, and adequately maintained, our research suggests

---

[14] Miller, Steven Robert. "PreCYdent: A New Search Engine Enters the Legal Research World." ALL-SIS Newsletter: Volume 27, Issue 3.
http://www.aallnet.org/sis/allsis/newsletter/27_3/PreCYdent.htm

[15] "copyleftist", "The Death of PreCYdent: Is Free Online Legal Research Sustainable?" Electronic Legal Research. http://elawresearch.wordpress.com/2011/02/02/the-death-of-precydent-is-free-online-legal-research-sustainable/ 2 Feb. 2011.

[16] http://web.archive.org/web/20100523204837/http://altlaw.org/

[17] http://code.google.com/p/google-ajax-apis/issues/detail?id=109

[18] To wit: Lois Law and Fastcase are lower-cost competitors with Westlaw and LexisNexis. Oyez and Justia are free but do not have a citator. LawPivot provides discount legal advice; LegalZoom provides templates for do-it-yourself legal services such as wills; Juridica is a risk management and financial hedging tool for law firms; Lex Machina scrapes intellectual property documents -- primarily patent filings -- to help law firms determine their clients' IP legal strategy.

that it will need to come from a non-profit and/or academic source, not the private sector, and require an external source of funding.

User Needs, Wants, and Expectations

So that we could familiarize ourselves with the UI and terminology of existing legal research tools, Professor Carver and the Berkeley Law School librarians gave us accounts on Westlaw, LexisNexis, and Bloomberg's legal research tools. Westlaw's search interface was terrifically crowded, with a separate search box for each type of search or filtering that you would want to use. Westlaw Next was significantly more modern in its interface, with only one search box, more descriptive labels, some AJAX functionality, and a better color scheme. We became semi-regular users of Westlaw Next over the course of the project as we needed to check our citation-finding results against an authoritative source.

While we did not do participant observation or user needs assessment for this project, we gained a great deal of insight from an interview with the law librarians at Berkeley Law School where we discussed which citator features lawyers relied on most and which they did not find useful. The librarians told us that lawyers often do not understand or trust the official six-category Shepardizing spectrum for whether a citation is positive or negative; no matter what the citator says, skeptical lawyers will read the case and judge for themselves. Thus, they held that a simple positive/negative indication on a citation would be good enough, and even that simpler version was not a high priority. Instead, the librarians emphasized the importance of depth of treatment, approximated by how many times a given case cites another, to determining the relevance of citing cases. Other features that they found useful included excerpts from the citing case to show *how* another case cites the one you're looking at; metadata tooltips on citation mouseover; the inclusion of legal statutes alongside case law (since new laws are sometimes passed specifically to overturn court decisions); and summary or headnotes at the top of the case.

The two main uses for a citator, according to the librarians, came down to two categories: "landmarks" and "landmines". Lawyers use citators as a springboard to find other cases that are more well-known and influential ("landmark") or more similar to the facts of the legal matter on which they are working. They also use citators to tell if a given case is a "landmine", i.e. that it has been overturned or called into question by newer court decisions, and is thus no longer a reliable basis for a legal argument. These insights shaped our case reader and citator designs and the order in which we decided to tackle citator features.

The Design

Service Design for a Free Legal Citator: The Berkeley National Reporter

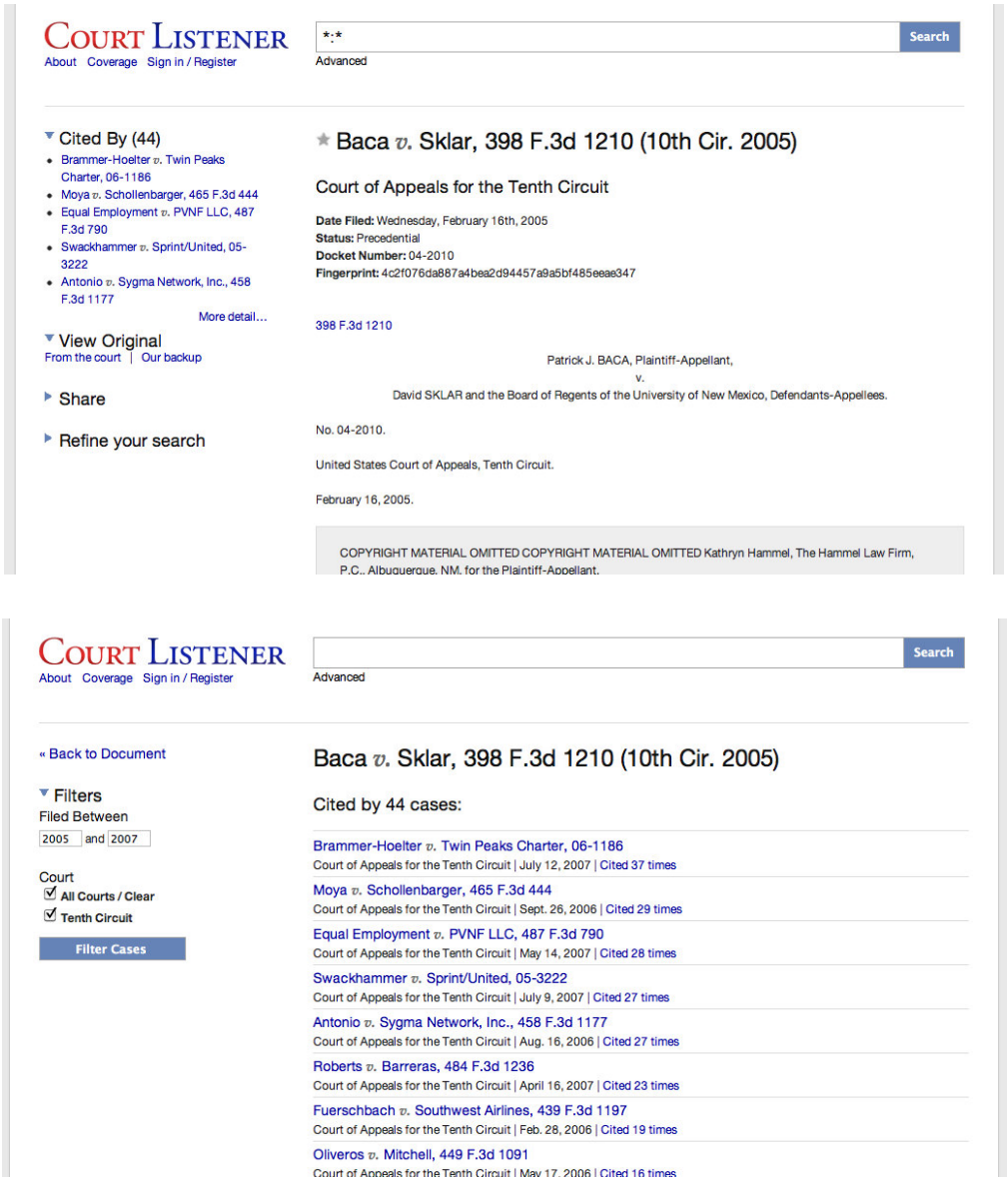
Based on our research, it seemed clear to us that the holy grail of making US case law and citation data accessible to all was not just a technical problem. So many startups and individuals had managed to write their own citator tools over the years, only to shut down or otherwise not meet this goal in other respects. Thus, we thought about what it would take for a public legal citator to survive and become a valued institution among lawyers and citizens.
It seems to us that three things would need to be in place for a free legal citator to succeed:

1. The citator service needs to be not-for-profit from the start and backed by a well-known law school. Raising adequate funds from both the host institution and outside philanthropic donors would be key.
2. The citator should be open source software from the very start. This mitigates two risks: first, it would make it futile for an incumbent like West to purchase the rival citator and shut it down or charge for access, and second, even if this iteration of the citator failed, others could pick up the code and build upon it instead of having to start from scratch.
3. The citator should keep it simple, stupid. Doing one thing well, instead of trying to do everything that Westlaw and LexisNexis do (poorly), is an important strategy in an environment of limited time and budget. As tech startup people say, all this service needs is the "minimum viable product" to get lawyers and others interested in using and get press.

Our project's advisor, Brian Carver, is in the process of creating an project along these lines, tentatively called the Berkeley National Reporter, based on CourtListener and our project's work. Already, representatives of UC Berkeley's I School and Law School, as well as open access advocate Carl Malamud of public.resource.org, have signed off on the idea of basing a free, public, open source citator service at the university. Although the eventual project may not be called "The Berkeley National Reporter" in the end, especially if other universities get involved, in our day-to-day work and in this paper we used this name to refer to the overall vision of a not-for-profit legal research institution.

Citator Design

After we got our citation finding and tracking code working on the live CourtListener.com website, we needed to add ways that the public could see the new citation information (besides the HTML links inside the case documents. Thus, we added a minimal citator front-end interface to the site.

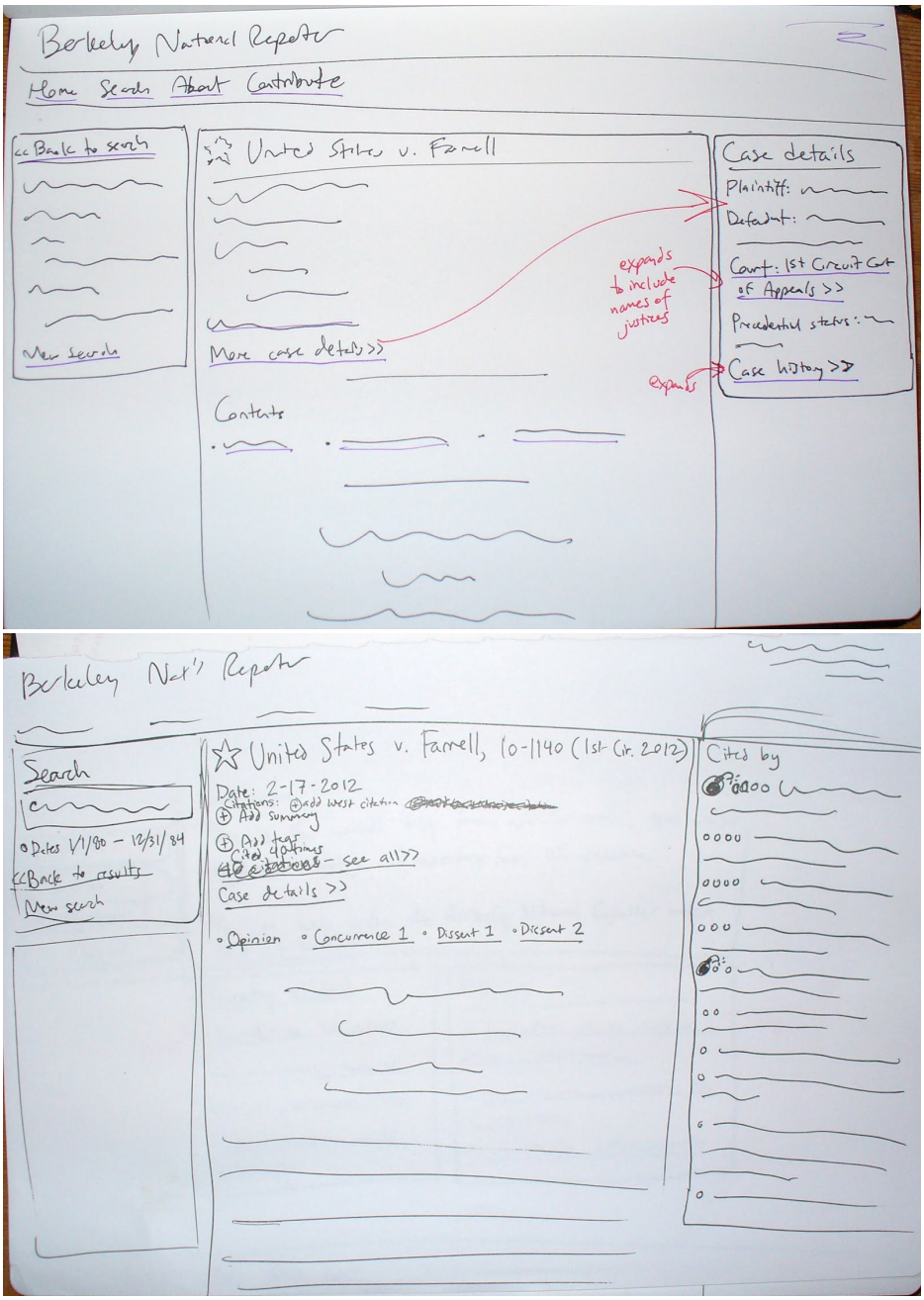


Case view and citations page on the live CourtListener website

The working citator on CourtListener.com is designed minimally primarily because CourtListener itself is primarily a legal feeds service and was not originally designed to be a site for case reading and in-depth research. Changing the focus of CourtListener to be a fully-featured legal research site, along the lines of the vision for the Berkeley National Reporter, would have been a far more radical change than either we or the CourtListener maintainer were willing to make in the time that we had.

We only included two filters on the CourtListener citator interface: by year and by jurisdiction. This matched the user needs expressed by the Berkeley law librarians and by our advisor, a practicing lawyer, that these were the most useful dimensions for us to support. This also made our filter sidebar not have too many elements to confuse users.

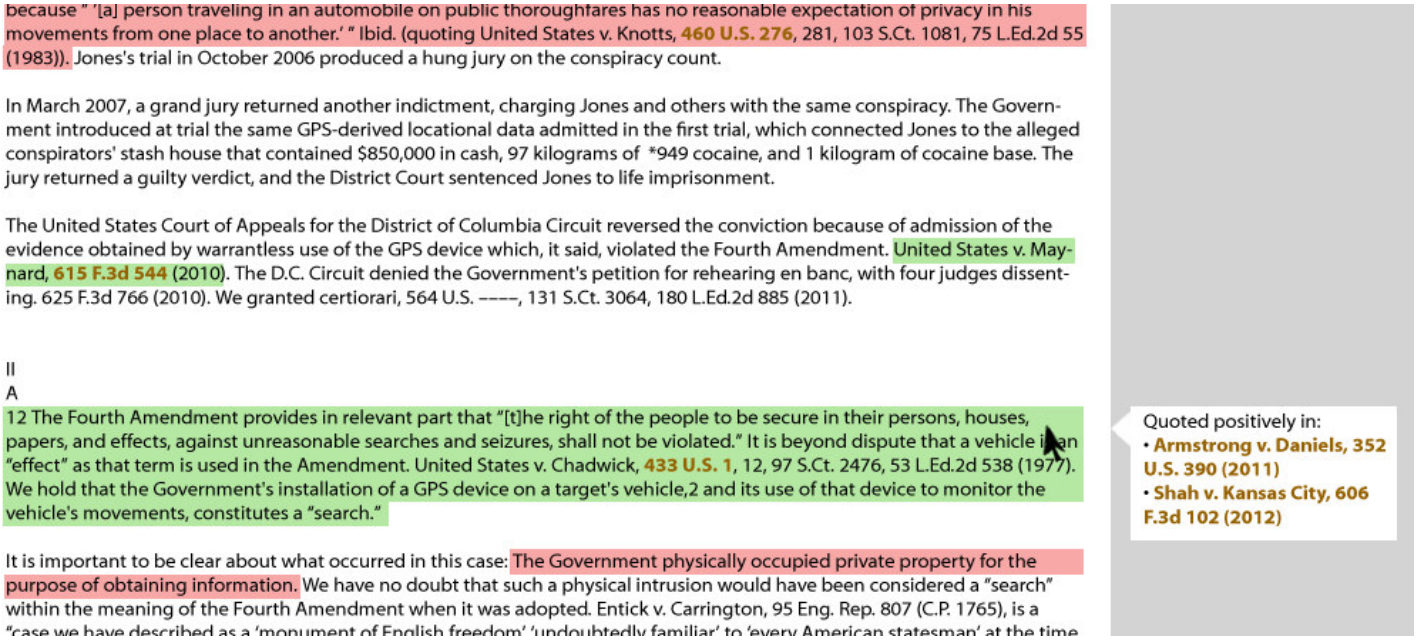
Although we did not create a separate Berkeley National Reporter website, we did make low-fidelity and high-fidelity mockups of what the Berkeley National Reporter citator interface might look like.



Low-fidelity case view mockups with case details pane and citation summary pane.

High-fidelity case view main page mockup

We attempted to organize case view pages such that users could access the information most important to their needs first, rather than presenting them with everything from the get-go and overwhelming them. As part of this, the site design has a right-hand sidebar where boxes with citation previews, additional case metadata, or annotations on the case text could appear. We also anticipated making it easy for users to jump to different opinions within the case text through anchor links. We also anticipated support for user-generated content and metadata, fetching placeholder case summaries automatically from the Wikipedia article for the case and adding an area for users to enter in the official West citation for a case if we did not already have it.

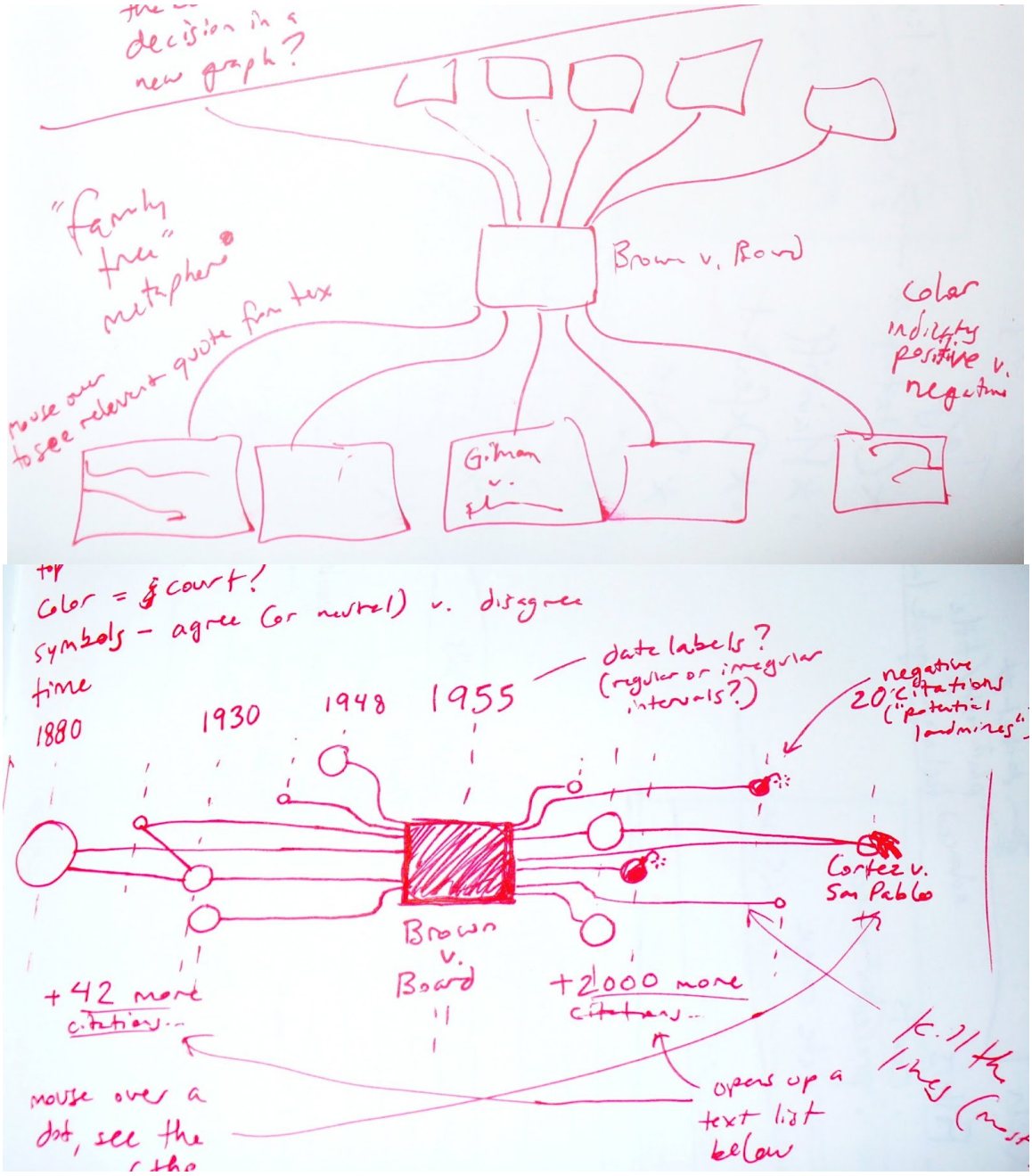


Line-by-line "freshness" interface mockup

We also thought about how the case reader interface could tease out individual "sinks" or legal ideas within a given decision and the different authority levels thereof. As in used law textbooks, passages of an opinion that had been quoted from by other decisions could be highlighted or otherwise emphasized. Based on quotations or page numbers, we could determine what part of a decision a given case was referring to; based on whether most of those citations were positive or negative (indicating whether or not those ideas were still considered good law), we could make the highlighting color green or red.

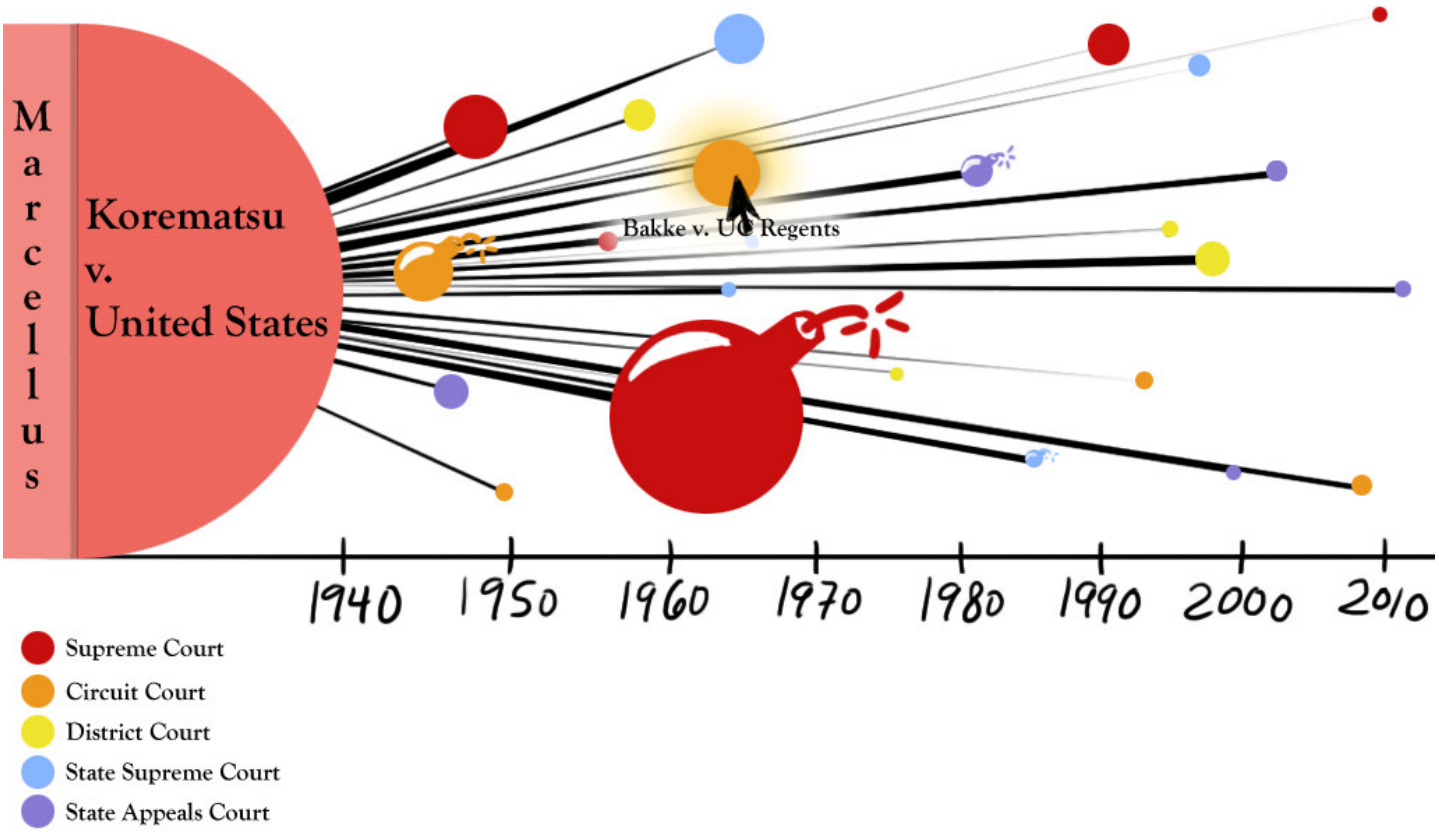
Visualizing the Case Law Family Tree

One of the key ideas of our project was improving people's access to the "family tree" of the law created by its citation network. Another of our design angles was building static and interactive information visualizations to make those relationships visible and understandable, especially by laypeople.



Paper sketches of ideas for visualizing legal citations

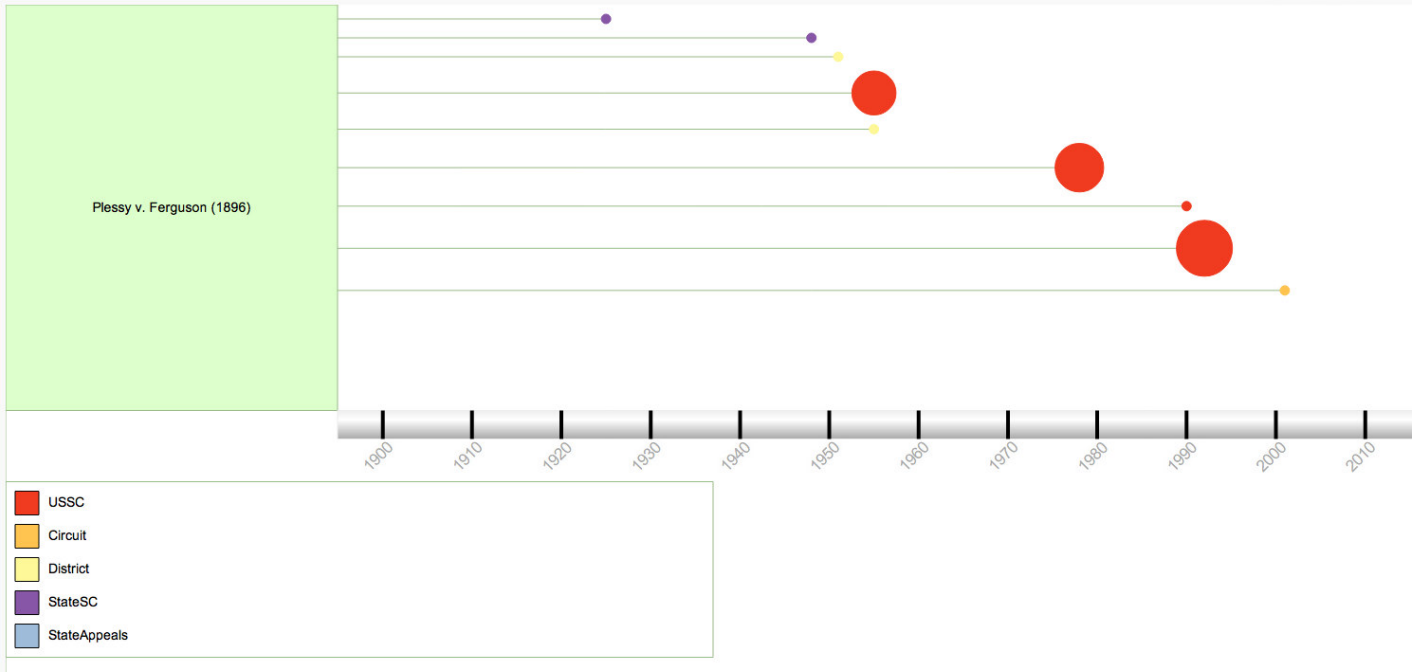
Our first ideas attempted to show both ancestors of a case (the cases that the given case cites) and descendants (cases that cite the given case). Eventually, we decided that the latter would be more interesting to users, since you can already see pretty well what cases a given case cites just by scanning the text and looking for the citation links.



Static digital sketch later on in the process

Our visualization ideas compressed several different data dimensions into a single image. Each node represents a case; color corresponds to different courts / jurisdictions; size of a node indicates its "influence" (how many cases cited that case); the position of a node on the x-axis indicates when the case was decided; and the node icon (a regular circle or a bomb) indicates if the citation is a negative one (and thus a "landmine" for the cited case).

In terms of interaction, we anticipated that mousing over a node would highlight it and show its title; clicking a node would animate the whole map, making the other citations disappear, the former parent node collapse into a rectangle, turn the clicked node into the lefthand parent hemisphere, and have the clicked node's citing cases spring outward to the appropriate places on the timeline.



Functional visualization prototype

We managed to prototype our visualization idea using real case data and real JavaScript code (using the graphics library Raphael). While we were not able to make the code-based prototype as slick and beautiful as our sketches before the filing of this report, all of the data dimensions are present.

Technical Implementation

Our citator was built as an addition to the existing code of CourtListener, so we used its technology stack, which includes LAMP (Linux, Apache, MySQL, Python) and Django for the main site, Solr for the search engine, and Celery[19] + RabbitMQ[20] for distributed task management.

Finding Citations

The first problem we had to solve was identifying citations within the text of a document. Working in our favour was the fact that legal citations have a predictable structure[21]. Working against us, however, were the formatting of the text from which we needed to extract citations and the vast variability of case names.

Currently the federal courts of the United States make their opinions available in PDF format. As a result, the majority of the documents in CourtListener have text that was as some point extracted from a PDF. Some of the challenges presented to us by the resulting text included:

- Extensive whitespace formatting
- Words split onto two lines with a hyphen in the middle
- Page numbers, line numbers, and footnotes included in the middle of the document contents
- Underlining converted to underscores

We received some very helpful advice about identifying citations from Itai Gurari, who has worked on legal search for Google Scholar. First, he suggested that we restrict our citation detection to a specific format and worry about variations later. The standard format of a basic legal citation is:

> *[volume] [reporter] [page], e.g. 410 U.S. 113*

For Supreme Court cases, this basic citation is followed by the year, in parentheses, e.g.

> *410 U.S. 113 (1973)*

For cases from the United States courts of appeals (circuit courts), the specific circuit is also included in the parentheses, e.g.

> *313 F.3d 500 (9th Cir. 2002)*

We took these two similar formats as our scope, ignoring other variations. Itai's second suggestion was to break the citation down into components and search in stages: first, look for the reporter portion of the citation, which can be matched against a list of known reporters; second, check for volume number immediately before; third, check for the page number immediately after. This takes care of the basic citation, and from there we can potentially identify additional information such as the court, date, and case name. The year and court can usually be located by looking for parenthetical expressions, but case names are much less predictable.

---

[19] http://celeryproject.org/
[20] http://www.rabbitmq.com/
[21] In fact, there is a whole book dedicated to explaining legal citation conventions.

A full in-text citation usually looks like: *Roe v. Wade, 410 U.S. 113 (1973).* A majority of case names have the form *Plaintiff v. Defendant*, which means that if we locate the 'v.', the text following it (and before the base citation) contains the name of the defendant. There are no restrictions, however, on the length or content of party names, which means that if we work backwards from the base citation (for the defendant) or the v. (for the plaintiff), there is no obvious guideline for how far back to look. We erred on the side of caution; for the plaintiff we only used the word immediately preceding the v. (if there was one), and for the defendant we stopped the search at 70 words, which is approximately the average case name length in the CourtListener database. In addition to the standard *versus* case name, we support the two forms where there are no opposing parties, which start with "In re" and "Ex parte," respectively. In summary, our process for finding citations implemented the following steps:

1. Tokenize the text of the document using the Natural Language Toolkit, nltk#, modified to treat federal case reporters as single tokens.
2. Identify base citations (volume, reporter, page)
3. Add as much additional information as possible, including the year, court, defendant name, and last word of the plaintiff's name.

Matching Citations

Once citations are identified, the next step is to match them against documents in the CourtListener database. A large portion of CourtListener's cases come from resource.org, and include official Westlaw citations (the base citation string discussed above). For many of these cases it is possible to find a 1:1 match, but some decisions are short enough that more than one can appear on the same page of a reporter, and therefore they share the same citation. The case name is necessary to find a definite match among such a set of decisions. Name matching, however, poses a particular challenge.

At the time of writing,, 17.5% of documents in CourtListener do not have formal citations in their metadata[22], and new documents which are added by court scrapers will also lack this information (as mentioned above, citations do not even exist when decisions are first published). For these cases, we attempt to match our extracted citations based on court, date filed, and partial case name.

Because case names are not standardized and may contain inconsistent abbreviations, doing exact phrase matches of our partial case name against the case names of the documents in the database had poor recall. To improve the coverage, we employed Solr's "minimum match" feature. Given a query string, one can specify a minimum number or percentage of its words which must be found in order to constitute a match. In order to match as much of the case name as possible, our search algorithm starts by requiring of the words to match, and decreases the minimum number until one or matching documents is found. This strategy, however, resulted in a large number of false positives. For example, if the partial case name is "Brown v. Board of Education," and the correct case is not in the database, the minimum match might return cases whose names contained just

[22] 132837 of 756270 documents (3 May 2012).

"Brown" and "Board" or "Board" and "Education." Even filtering by court and date filed, erroneous matches occur. To avoid these mis-matches, we performed a verification query for the full case name of the purported match in the text of the original citing document. The resulting algorithm for matching citations to documents in the database can be summarized as follows:

1. Construct a query filter for date filed and court, if available.
2. Search the database for a document with a West citation field that exactly matches the base citation as well as the filter parameters.
3. If an exact citation match is not found, search for documents whose case name matches the partial case name identified in the text.
4. Verify possible name matches by doing a reverse search for the case name of the matched document in the text of the original citing case.

The completed process for adding citation information to a document in CourtListener's database is:

1. Identify all citations within a document's text.
2. For each citation, look for a matching document in the database.
3. If a match is found, add it to the original document's list of cases cited, and update the incoming citation count on the matched document.
4. Annotate the citations in the document with HTML, including links to matched documents or a placeholder information page.

The resulting script can be run as a standalone over a subset of documents in the database, and it is also called whenever a new case is added to CourtListener. With over 700,000 existing documents at the time of writing, adding citations was a substantial computation. To minimize the performance hit to the CourtListener as much as possible, we had to optimize and limit the number of queries to the Solr index. The job also took advantage of the Celery queue to distribute tasks.

Results and Statistics

When run over the database, our code created 4.2 million citation links between documents.[23] The total number of citations identified was even higher—though we don't have an actual count—because it included federal reporters which are not yet included in CourtListener. A sample run over 50 federal appellate decisions found 234 citations and matched 181 of them, a 77% success rate. Supreme Court opinions fared even better, with 93% of citations successfully matched to documents in the database. This difference is most likely due to the fact that the Supreme Court heavily cites its own past decisions, and because CourtListener has excellent Supreme Court coverage.

According to our citator, the two cases on CourtListener with the most incoming citations are:

1. Strickland v. Washington, 466 U.S. 668 (1984) has 9,592 cites. As a case about the standard for establishing ineffective assistance of counsel, it is routinely cited in habeas matters.
2. Anders v. California, 386 U.S. 738 (1967) comes next with 8,948 cites, an important case about the right to counsel on appeal of indigent defendants.

[23] And executed over 10.9 million Solr queries!

Future Work

While our project tackled the most fundamental part of building a citator—building a working citation identification and tracking tool—for the citator to meet users' needs (as identified by the UC Berkeley law librarians and others) there are a number of features that would need to be added. Items on the to-do list for our citator include:

- Greater coverage of the citations we identify: Currently, our citation identification tool can only identify federal court citations, not those of state, district, or local courts.
- Depth of treatment tracking: "Depth of treatment" refers to the degree to which a case discusses another case--whether it is included in passing or extensively analyzed. A simple way to calculate depth of treatment scores for citations in an automated fashion would be to track how many times a case cites another case. This would require that the algorithm be able to identify subsequent citations in a document which have a different, truncated format (or even just "id.")
- Natural language processing and/or sentiment analysis for shepardizing dimensions: In order to generate a "freshness score" or even a simple thumbs up/thumbs down guess as to whether or not a case is still authoritative law, one would need to create an algorithm evaluating the text around a citation to guess whether the case agrees or disagrees with the case being cited. While our team had interest in this area of the citator problem, we did not have the time or resources to pursue it as part of this project.
- Bayesian algorithms and/or human review for citation finding: In spam filters, a Bayesian algorithm evaluates an email and, based on known past data, classifies it as spam or "ham" (desired email). Such algorithms can be used in other applications as well. One could imagine, instead of hard-coding a set citation-finding algorithm, "training" a computer to tell the difference between a citation and non-citation text using a corpus of case documents with the citations already identified and an interface for human review of the citation finder's decisions. It is quite possible that such an algorithm would eventually be more accurate than our current citation-finding algorithm, especially at tricky problems like identifying where the plaintiff name in a citation begins. However, we did not attempt this approach to find out its relative efficacy.

Besides the citator tool in isolation, there is also much work yet to be done in making a free, not-for-profit national reporter initiative as a whole a reality:

- Onboarding open source software contributors: While CourtListener is already open source, technical changes to CourtListener's technology suite would make it much easier for new volunteers to get involved. Extracting tricky dependencies such as Solr and Celery (so that you would only need to install and operate those libraries if you happened to work on the parts of CourtListener that interact with them), enabling development installations of the site to work from sqlite instead of heavier-weight MySQL, making CourtListener's setup instructions use a virtual environment by default to sandbox its dependencies, or switching version control systems from mercurial to the more popular git all would reduce barriers to entry and make it more likely that others would work on CourtListener's code base.

- User-generated content and metadata: CourtListener currently does not leverage crowdsourcing to fill in gaps in its case law database or metadata about cases. Changing the site to make it easy for interested volunteers to fill in missing information--while providing safeguards against vandalism and checks for quality--is highly recommended because quite a lot of case law metadata--West citations, case summaries, legal topic tagsnomies--is difficult or simply impossible to extract by automated means. The case of Wikipedia is notable here: although Wikipedia is not targeted towards the legal community, it has nonetheless attracted volunteers to crowdsource articles on many notable cases, including accurate metadata and even West citations. If the Berkeley National Reporter gained sufficient momentum, it is not unreasonable to believe that, for instance, law students would contribute summaries of cases based on briefs written for class, or more experienced lawyers would correct errors in passing as they conduct their research. Creating an interface and backend architecture to facilitate that involvement would be a necessary prerequisite.
- User experience evaluation and iteration: Finally, while we attempted to follow the recommendations and needs expressed in our discussions with law librarians, lawyer friends, and our project advisor, we did not conduct a full user needs evaluation or test our legal research user interface vision. Qualitative and quantitative evaluation of CourtListener's (and, eventually, the Berkeley National Reporter's) user experience with both lawyers and non-lawyer users would be an essential step in ensuring that the project meets its goal of making the law clear and accessible to all while still being an effective legal research tool.

Finally, given this map of the relations between cases, academics could explore any number of citation analysis-based research questions. For instance, not even paid subscriber access to Westlaw will let you answer the question of what courts or cases are the most influential or, in a brute-force sense of the question, are cited the most.[24] The calculated influence of cases could be used to improve CourtListener's search results, similar to Google's use of PageRank.[25] Technical improvements, such as creating a separate API for citation data and/or changing the database backend for the citations-tracking part of the site from a SQL database to a node-based database architecture (reflecting the networked nature of the data), would further assist such research.

Conclusion

Although there are many improvements that could be made to our legal citator, the fundamental citation-finding and tracking functionality already opens up many possibilities. We are hopeful, given CourtListener's relative longevity, the open source (and thus reusable and patchable) nature of the code, and exciting plans for public legal citator support at UC Berkeley or other institutions, that this potential will be realized.

---

[24] A superficial look at such a question finds that all but one of the top twelve most cited cases are Supreme Court cases, unsurprisingly; most seem to deal with basic civil procedure matters.

[25] Credit goes to Brian Carver and Michael Lissner for discussion of some of these possibilities.

Acknowledgements

We'd like to thank our adviser, Brian Carver, who was closely involved in the project; Michael Lissner, for his hours of patient assistance and work integrating our code into CourtListener; Itai Gurari, for his advice on implementing a citation-finding algorithm; and the Berkeley Law School librarians, for taking the time to meet with us and share their experiences with legal research.