MIT Computational Law Report •

Tutorial: Citation Analysis via Caselaw Access Project API

Andy Gu

Published on: Dec 07, 2021

License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)

At the Library Innovation Lab at Harvard Law School, we often think about directions for empirical studies of U.S. case law. Through the Caselaw Access Project (CAP), we maintain and operate a corpus of over five million cases and forty-six million case citations that are digitized and structured from bound volumes owned by the Harvard Law School Library. Via an open-source API, programmers can access case metadata and full-text. Additionally, CAP allows users to download all citations between cases in the above corpus. A bulk list of supported reporters may be found on the project website.

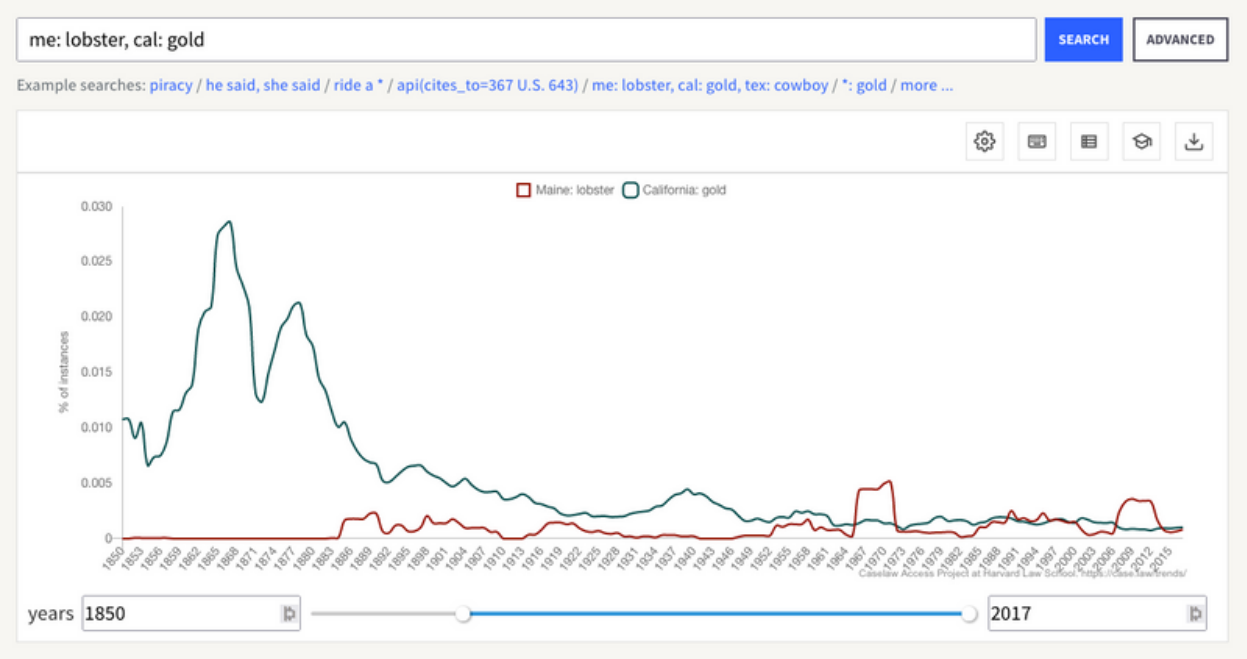
This corpus of case and citation data permits users to deeply explore citation patterns in the law. Researchers have used case citation analysis in myriad ways to understand patterns in U.S. case law, sometimes as a proxy for case influence[1] or as a partial measure of judicial quality.[2] For instance, the Citing Slavery project by Justin Simard uses case and citation data to surface the legacy of slave cases through citation.[3]

As stewards of the Caselaw Access Project, we are interested in extending features that empower users to quickly understand citation and case data. In particular, we are curious about aiding approaches to exploratory data analysis (EDA), an approach by which users often utilize visualization and analysis tools to retrieve the general characteristics and relevant patterns of an underlying dataset.

In order to ease EDA over the Caselaw Access Project's dataset, we have extended the project's visualization tools in order to allow users to build fast, flexible time series data about cases and case citations. This post will demonstrate a new feature that quickly builds complex timelines of case citations, walks through a number of use cases for researchers, and highlights opportunities for future direction.

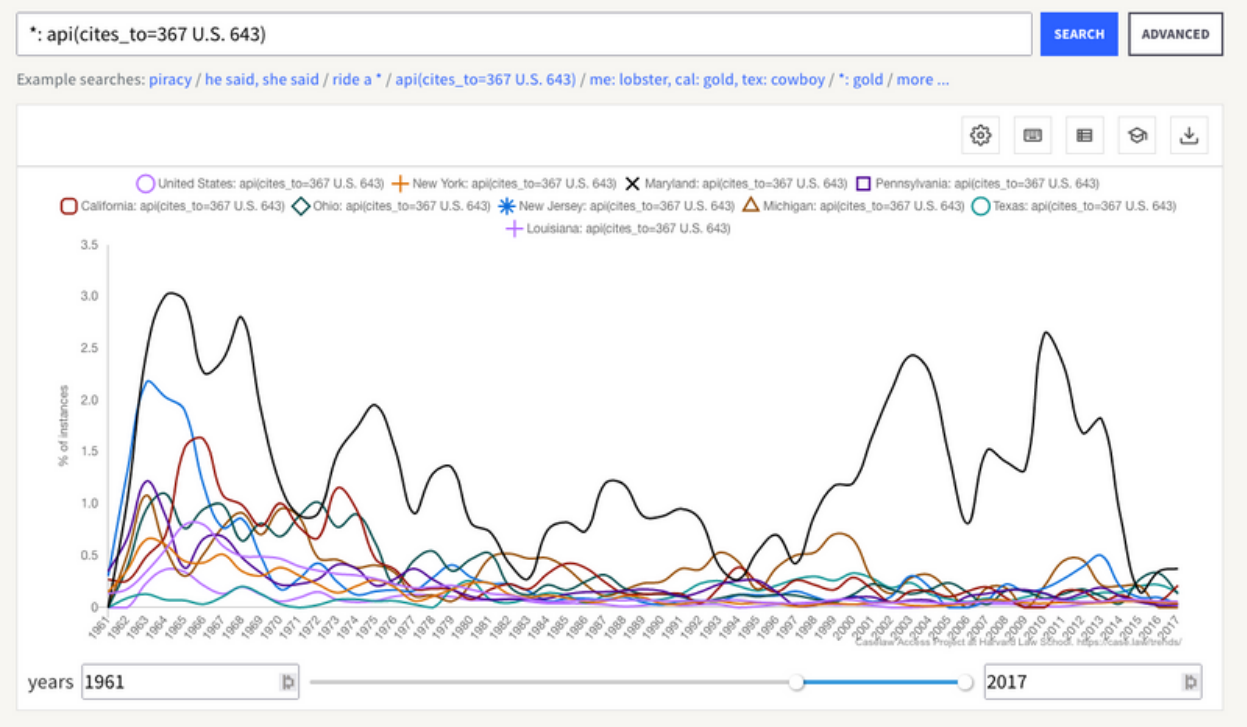
Programmatic Timeline Generation

Previously, the Caselaw Access Project's Historical Trends tool supported word or word phrase queries of up to three words. Users could query for the frequency of the instances of word phrases or the frequency of the cases in which those phrases appeared. Additionally, users could restrict their searches only to instances where the word phrases appeared in a given jurisdiction. For instance, the following graph shows the relative frequency of the term "lobster" in Maine and the term "gold" in California as a percentage of all tokens in their respective jurisdictions:



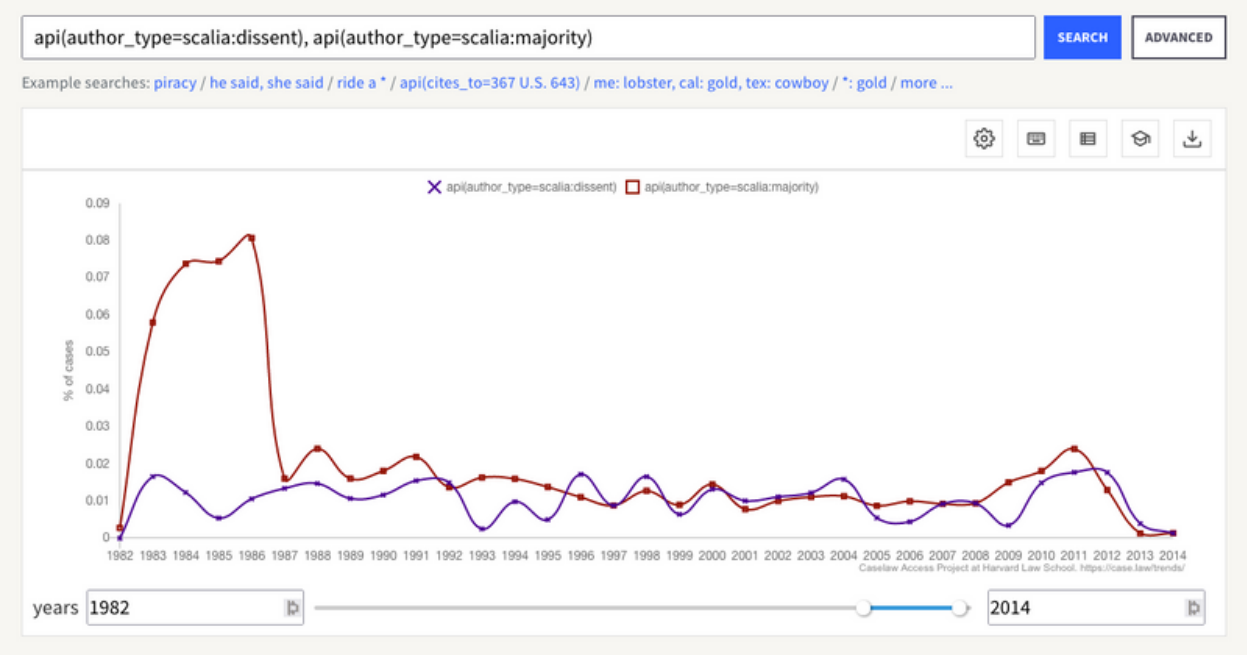
Query: me: lobster, cal: gold

We have extended this tool to also support more granular queries of cases using the Caselaw Access Project's Cases API. Using the `api()` tag, users can supply any parameter compatible with the Cases API. The Trends Viewer will then retrieve all matching cases, aggregate them by date (and also optionally by jurisdiction), and render a timeline of the resultant data. For instance, a user can display the top ten jurisdictions that make decisions citing Mapp v. Ohio over time with the following query:



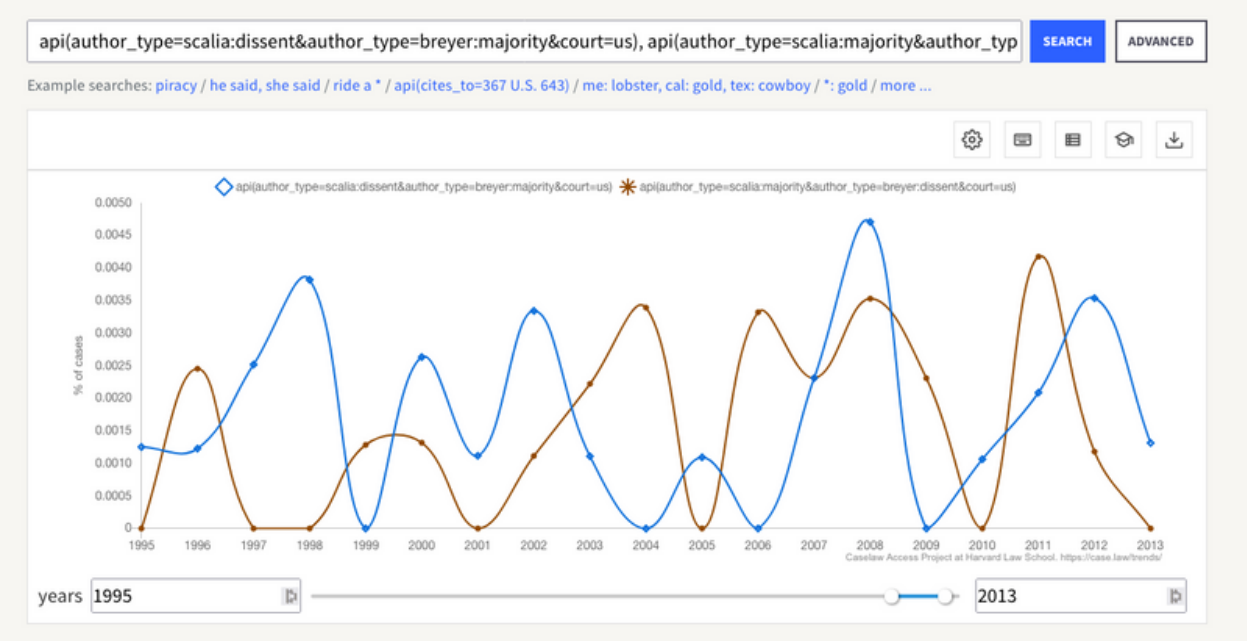
Query: api(cites_to=367 U.S. 643)

Due to the large number of parameters supported by the [Cases API], users can generate powerful timelines that expose interesting relationships in case law. Generating these timelines using the Trends tool is much faster than manually downloading the corpus and parsing it with a scripting language. For instance, if a user wanted to map dissenting versus majority opinions by Justice Scalia, they could produce the following query:



> **Query: api(author_type=scalia:dissent), api(author_type=scalia:majority)**

The number of possible relationships that users can interrogate increases with each additional parameter added to the search. By adding additional parameters, one could map cases in which Justice Scalia wrote a dissenting opinion and Justice Breyer wrote a majority opinion on the United States Supreme Court with cases in which Justice Breyer wrote the dissenting opinion and Justice Scalia wrote the majority opinion:

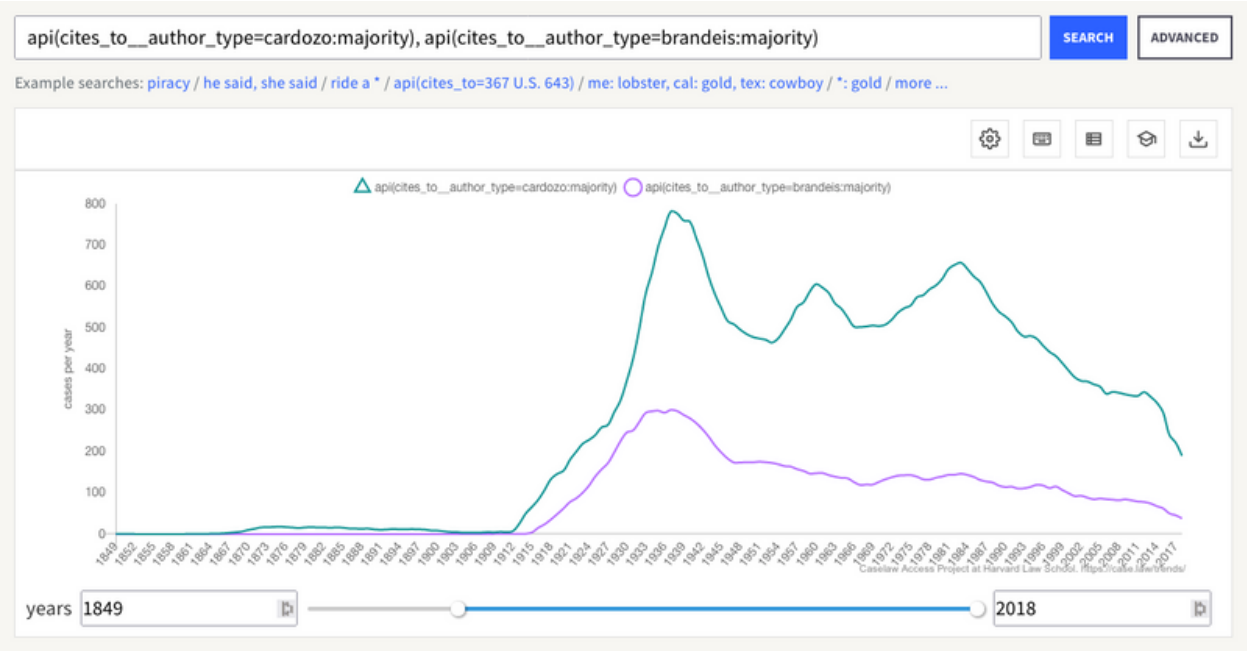


> **Query:**
> **api(author_type=scalia:dissent&author_type=breyer:majority&court=us),**
> **api(author_type=scalia:majority&author_type=breyer:dissent&court=us)**

Second-Order Citation Queries

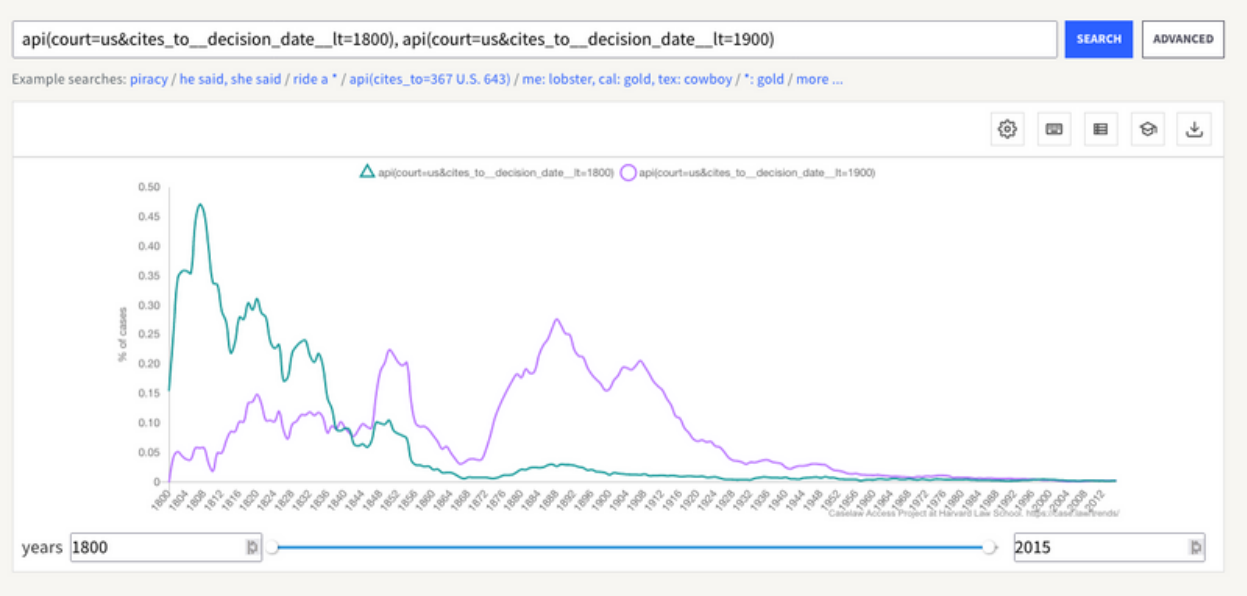
In addition to filtering cases by particular attributes, we also extended support for users to query cases by the types of cases they cite to. For instance, by prepending the `cites_to__` prefix to any parameter accepted by the Cases API, users can query for cases that cite to cases matching those parameters.

A sample use case involves measuring citations to a case as a proxy for the influence of the case author. In the following query, users can query for the number of citations to a case whose majority opinion was written by Justice Cardozo with cases whose majority opinion was written by Justice Brandeis.



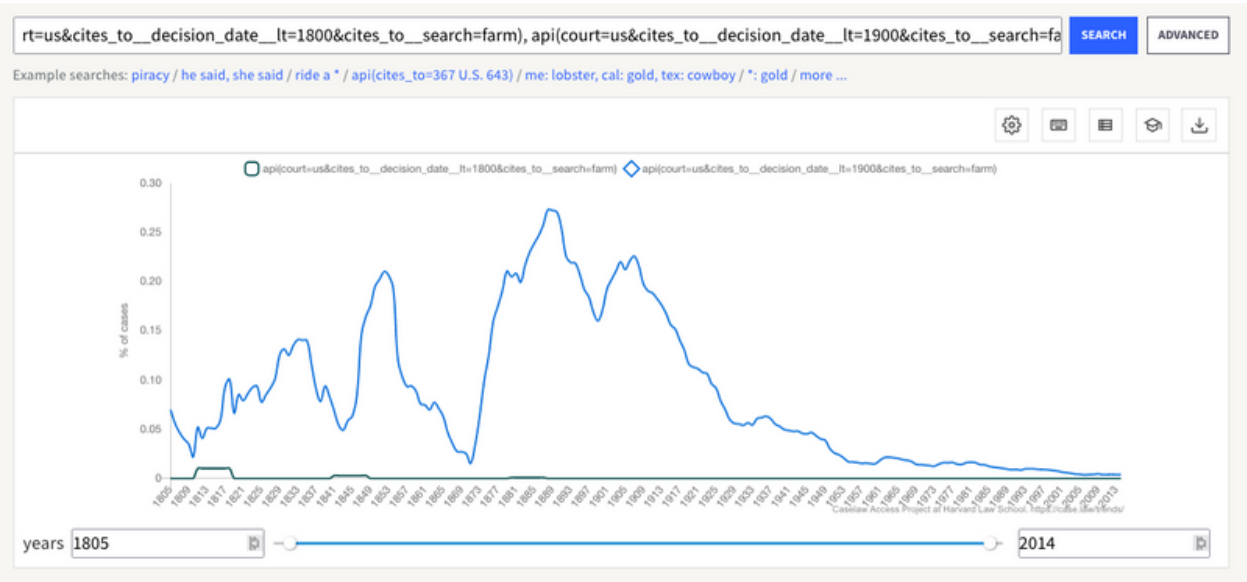
> **Query: api(cites_to__author_type=cardozo:majority),**
> **api(cites_to__author_type=brandeis:majority)**

Users can prepend `cites_to__` to any field used for filtering data, such as jurisdiction, decision date, or search. In the following example, users can build a timeline of cases written by the Supreme Court that cite to cases decided prior to 1800 and 1900, respectively:



> **Query: api(court=us&cites_to__decision_date__lt=1800), api(court=us&cites_to__decision_date__lt=1900)**

Like before, these parameters can be combined with any other parameter compatible with the Caselaw Access Project's Cases API, further increasing the number of possible relationships one can surface in case law. Taking the previous example, users could further filter the above timeline's cited cases to also only include cases mentioning the term "farm."



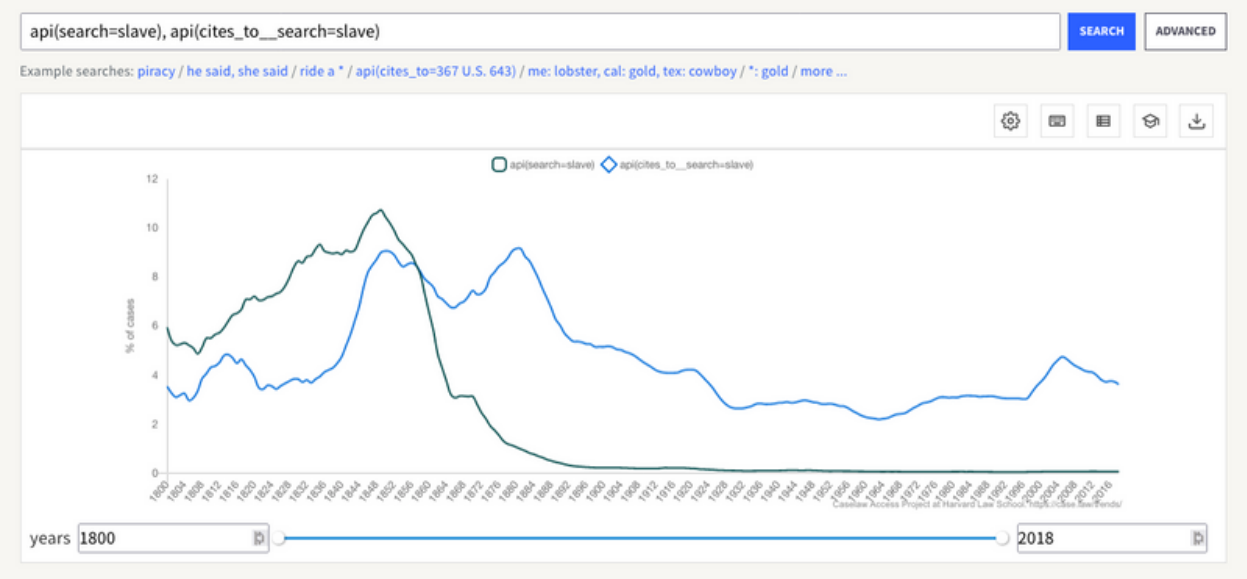
> **Query:**
> **api(court=us&cites_to__decision_date__lt=1800&cites_to__search=farm),**
> **api(court=us&cites_to__decision_date__lt=1900&cites_to__search=farm)**

In order to minimize load on the cluster given our current data schema, the `cites_to__` feature will look for a maximum of 20,000 cases to cite to. If the set is greater than that, 20,000 cases will be chosen randomly to cite to. This may generate misleading information if the actual size of the cited cases is extremely large, such as the set of all cases in California. In future work, we hope to restructure our underlying backend to support matching against all relevant cases, pending user feedback.

Case Study: Citing Slavery

The ability to query case trends not only by their characteristics but also what they cite to dramatically expands users' power when conducting EDA. Taking an example using previous citation research, we considered Justin Simard's exploration of the legacy of slave cases through their citations in American case law.[4] To find slave cases, Simard used electronic databases like Westlaw and LexisNexis to find opinions discussing slavery prior to the ratification of the Thirteenth Amendment, then queried for citations to them between 1985–2020.

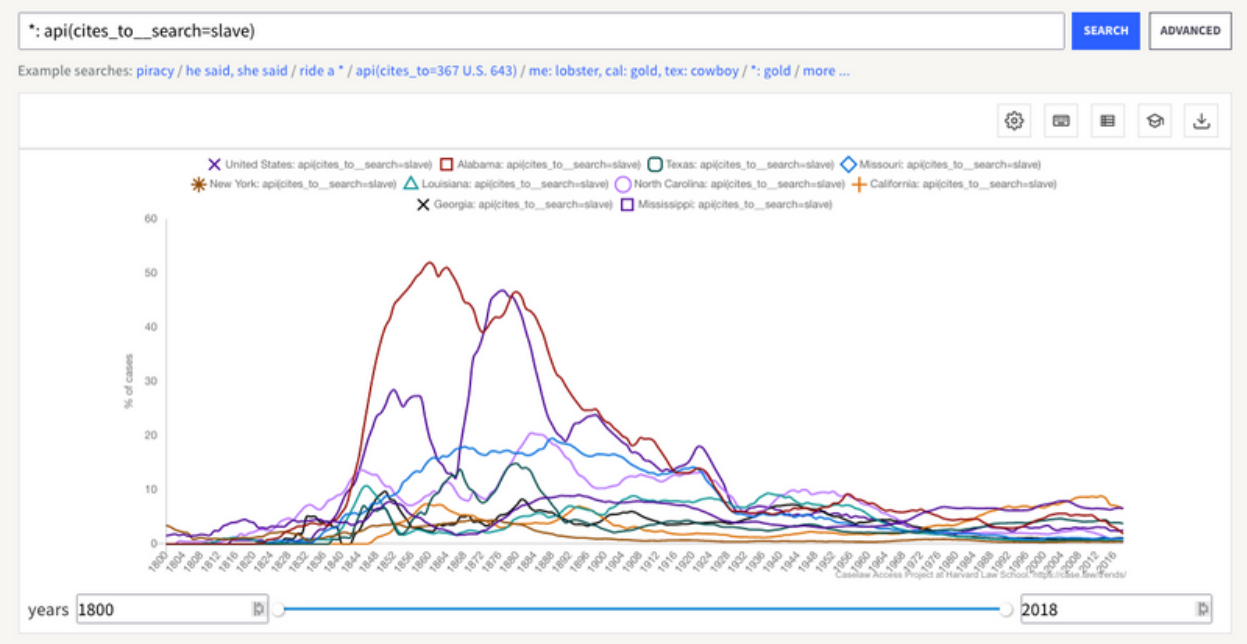
While this individual analysis is ultimately necessary to support a research hypothesis, the tools introduced here would allow a researcher in Simard's position to much more quickly evaluate his hypothesis before investing a more significant individual effort. For instance, the following query displays the percentage of cases that contain the word "slave", compared to the percentage of cases that cite another case containing the word "slave."



> **Query: api(search=slave), api(cites_to__search=slave)**

The visualization hints towards a broad theme in Simard's work: despite a large drop in cases mentioning slavery (a rough proxy for slave cases) after the Civil War, those cases likely live on as legal precedent. Certainty in these results is clearly diminished due to the removal of several layers of detail — unlike Simard's research, this query does not account for terms where the term "slavery" is used as an analogy our outside of the context of American chattel slavery.

However, the tool's power in exploratory data analysis dramatically increases options for researchers with an intuition about case law that would otherwise be difficult to confirm. One can imagine a researcher extending the above search and asking whether different jurisdictions or courts are more likely to cite to slave cases:

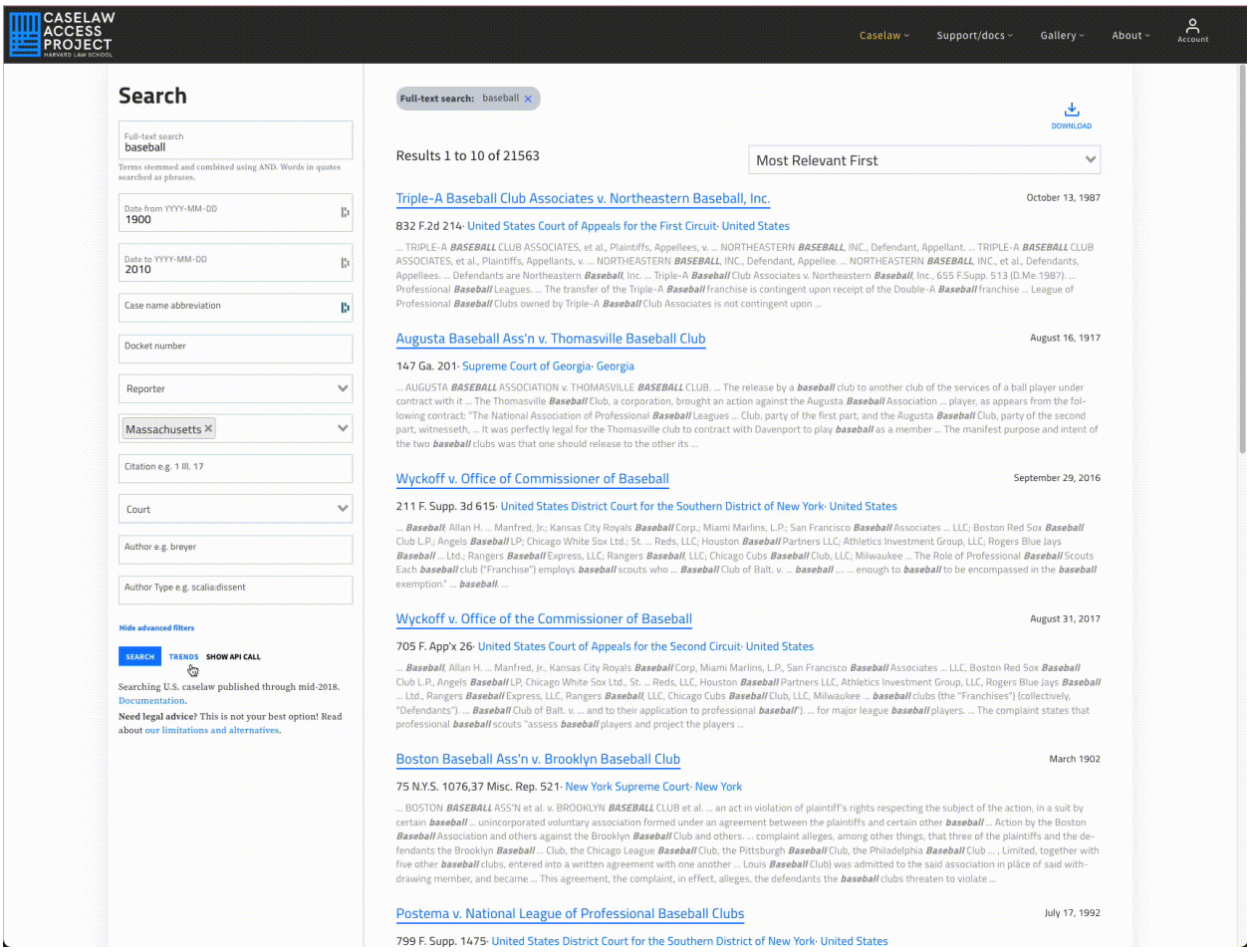


> **Query: *: api(cites_to__search=slave)**

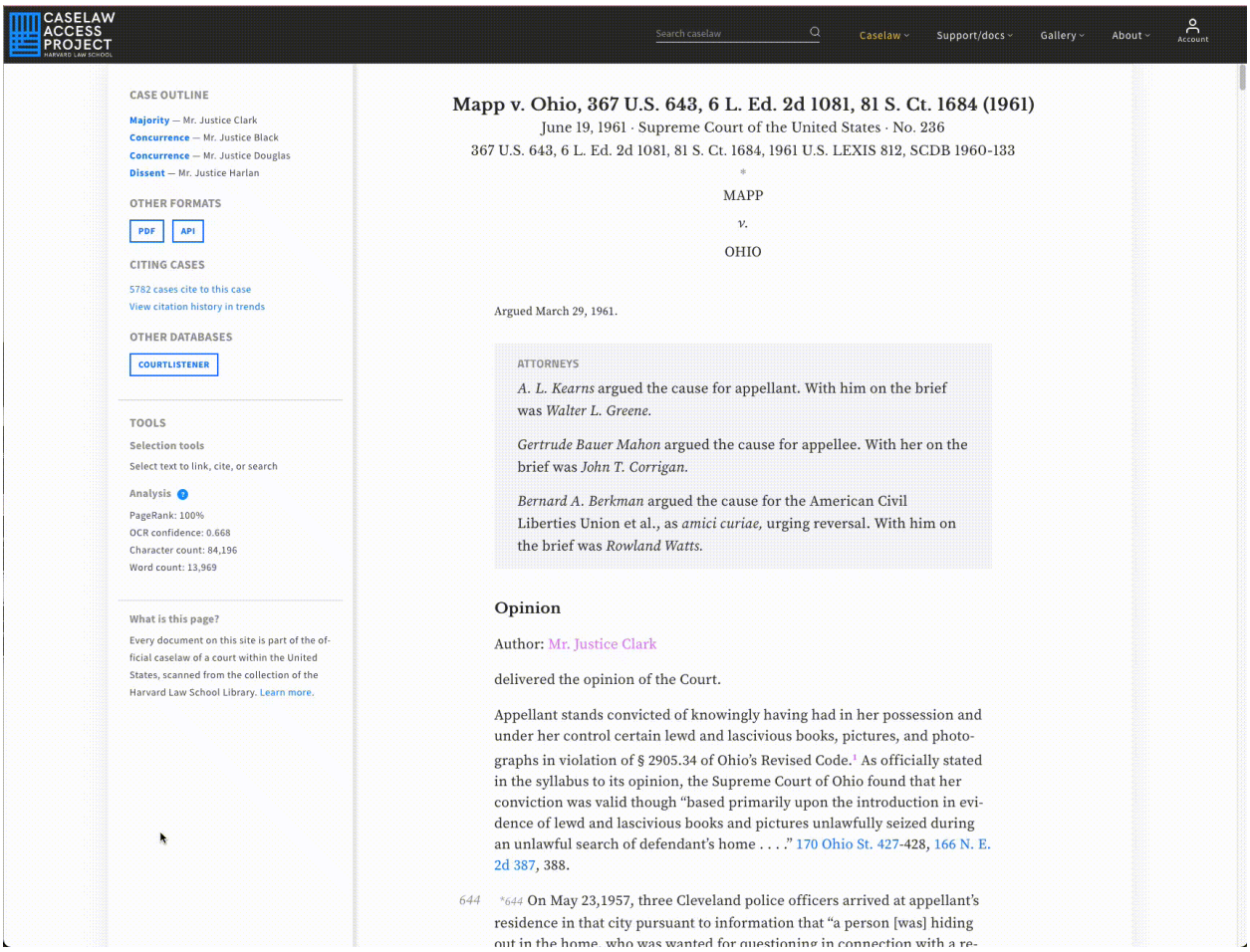
While there are certainly limitations to query flexibility in the Trends Viewer's current iteration, its flexibility and relative ease of use make EDA over case law easier. Given an intuition about the law, a researcher does not need the technical capability to download and process a citation graph and case metadata in order to understand a pattern of cases. Using these visualization tools, they can better understand potentially significant patterns in case law earlier in the research process.

Search Interface Changes

Due to the complexity of the API, researchers may find writing queries to be more difficult, especially early on. To make usage easier, we have built a conversion tool from which users can convert a search on the Caselaw Access Project's Search Tool into a Trends Query. From a search, users can click a Trends button that turns an existing search into a timeline. A sample video of usage is shown below:



Users can also find the citation history for a particular case by clicking a “View citation history in trends” button:



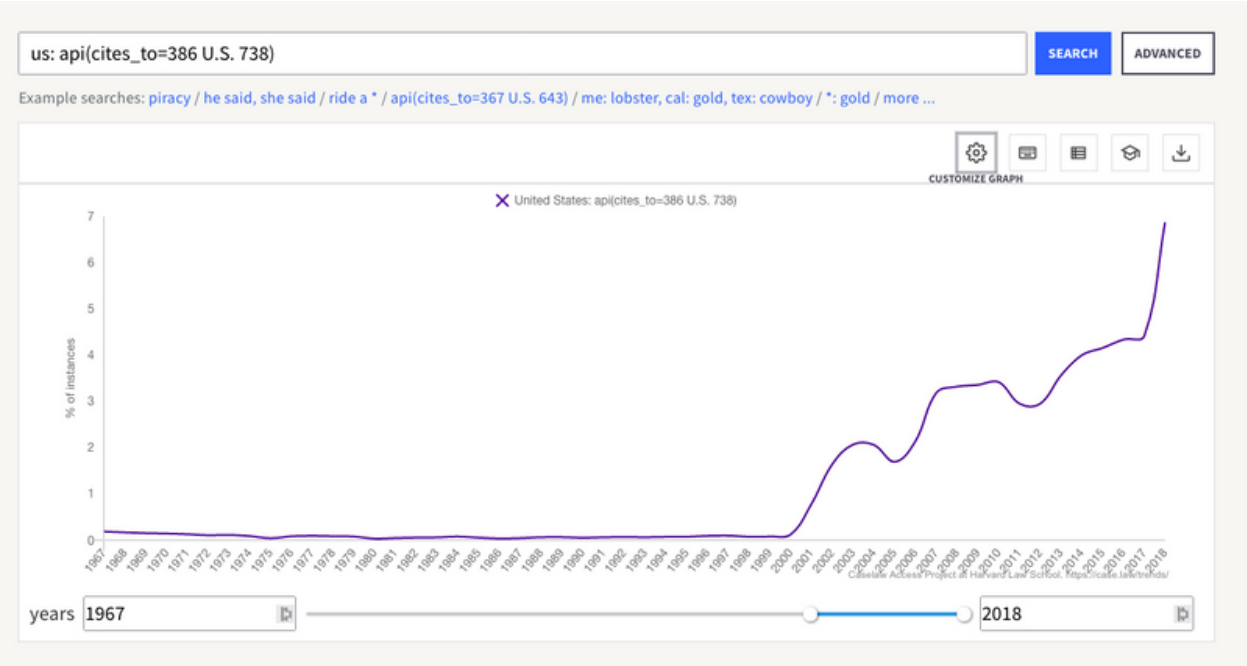
Future Direction and Limitations

While the Trends Viewer empowers users to conduct powerful searches of case law, we plan on expanding our work to surface new relationships and make searching easier. For instance, our extensions are currently limited in granularity of aggregations. While users can retrieve data parsed by year, jurisdiction, or a combination thereof, they cannot aggregate data along other facets such as by court, judge, or reporter. We hope to build support for more flexible aggregations in the future.

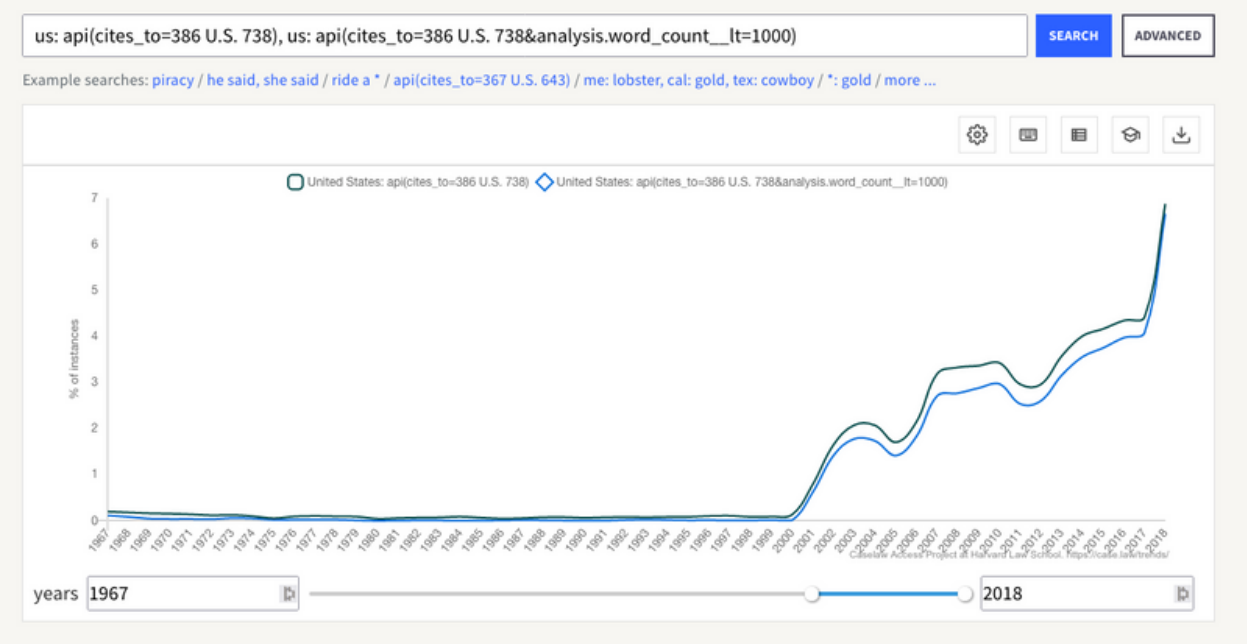
Additionally, we hope to increase the maturity of our search-based assessments in the future so that users with less technical backgrounds can more easily navigate the search interface. For non-technical users or students looking to explore case law, more complete guidelines around how to navigate patterns in the CAP dataset could increase future adoption. These goals may be substantiated through a more intuitive search interface that automatically generates a corresponding timeline.

Importantly, the accuracy of these tools is limited by the accuracy of the underlying dataset. Noise can be introduced both in the digitization process and in the data itself. The Caselaw Access Project structures case text using a combination of OCR tools and human review, then extracts citations using the eyecite library developed in collaboration with the Free Law Project. While the Library Innovation Lab takes great pains to improve accuracy, automated digitization tools at scale inevitably introduce errors. Citation extractors also may exclude interesting context such as signals or preceding words, which further color the context around a particular citation.

Our results can also be influenced by the content of the underlying data, introducing certain biases when conducting aggregate analyses of the law. For instance, take the following graph navigating citations of Anders v. California over time.



It seems puzzling that citations to Anders v. California drastically increased since the year 2000. However, this phenomenon is more likely attributed not to some substantive change in criminal procedure but to the publication of the Federal Appendix, which began in 2001 and covers federal appellate decisions not expressly designated for publication in the Federal Reporter. The next graph hints that the bulk of these cases are under a thousand words and possibly represent brief decisions responding to the filing of an Anders brief.



It is important to recognize that the integrity of the visualization is controlled by the integrity of the underlying dataset. While the Library Innovation Lab takes care to ensure the accuracy and completeness of its data, researchers should take care to download and evaluate the shape of the underlying case data for their own research goals.

Conclusion

We believe that citation analysis can surface new perspectives to our understanding of case law, and we are excited to see how users will leverage these tools to expose legal trends. With the Trends tools, users can quickly replicate historical citation research as well as conduct exploratory analysis of the law with minimal technical infrastructure. We hope that the use cases outlined above will provide users with the insights necessary to conduct independent discoveries in American case law. If you have any ideas for how we can improve these tools for users, please do not hesitate to reach out to us at info@case.law.

Header image generated with Wombo

Footnotes

1. William M. Landes, Lawrence Lessig & Michael E. Solimine, Judicial Influence: A Citation Analysis of Federal Courts of Appeals Judges, 27 J. LEGAL Stud. 271 (1998). ↩

2. Stephen J. Choi & G. Mitu Gulati, Choosing the Next Supreme Court Justice: An Empirical Ranking of Judge Performance, 78 S. Cal. L. Rev. 23 (2004). ↩

3. Justin Simard, “Citing Slavery” 72 Stan. L. Rev. 79 (2020). Available at: https://review.law.stanford.edu/wp-content/uploads/sites/3/2020/01/Simard-72-Stan.-L.-Rev.-79.pdf ↩

4. *Id.* ↩