

*Legal Citation Analysis with **CourtListener & Cobaltmetrics***

Luc Boruta & Damien Vannson — Thunken
luc@thunken.com — @thunkenizer
6:AM Altmetrics Conference, 2019/10/09

<http://gph.is/X18Wen>



THUNKEN

Acknowledgements

COURT LISTENER



Cobaltmetrics

- **Mike Lissner**, Free Law Project
- **Casey Scott McKay**, Thunken

Cobaltmetrics: web-scale citation tracking

Metrics are a sampling game.

Imbalanced datasets reinforce discrimination.

It is not up to citation aggregators to decide what is citable,
our role is to **observe all citation patterns on the web.**

Cobaltmetrics: web-scale citation tracking

Cobaltmetrics crawls the web to index
hyperlinks and PIDs as first-class citations.

The web is our corpus, and our URI transmutation API
collates citations to all known versions of a document.

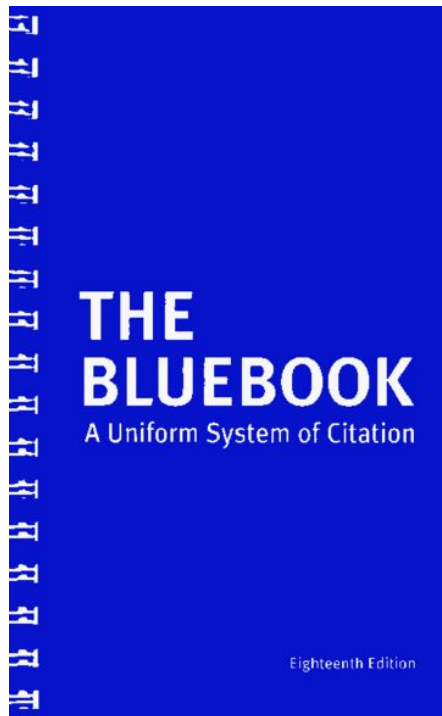
Cobaltmetrics: web-scale citation tracking

- Wikimedia: all projects, all languages
- StackExchange: all projects, all languages
- **US legal opinions, via CourtListener**
- CommonCrawl
- Hypothes.is annotations
- Usenet posts, via the Internet Archive

Legal citation analysis

- Scope: US legal opinions from 403 jurisdictions
- Data source: CourtListener
- Citation extraction:
 - Opinion to opinion: done by CourtListener
 - Opinion to anything else: done by Cobaltmetrics

Challenges: overcomplex citation rules



B1 Structure of Legal Citations

- B1.1 Citation Sentences and Clauses
- B1.2 Introductory Signals
- B1.3 Explanatory Parentheticals

B2 Typeface for Court Documents

B3 Subdivisions

B4 Short Citation Forms

B5 Quotations

- B5.1 Generally
- B5.2 Block Quotations

B6 Abbreviations, Numerals, and Symbols

B7 Italicization for Style and in Unique Circumstances

B8 Capitalization

B9 Titles of Judges

B10 Cases

- B10.1 Full Citation
 - B10.1.1 Case Name
 - B10.1.2 Reporters and Pinpoint Citations
 - B10.1.3 Court and Year of Decision
 - B10.1.4 Pending and Unreported Cases
 - B10.1.5 Weight of Authority and Explanatory Parentheticals
 - B10.1.6 Prior or Subsequent History
- B10.2 Short Form Citation

B11 Constitutions

B12 Statutes, Rules, and Restatements

- B12.1 Full Citation
 - B12.1.1 Federal Statutes
 - B12.1.2 State Statutes
 - B12.1.3 Procedural Rules, Restatements, Uniform Acts, and Similar Materials

B12.1.4 Federal Tax Materials

B12.2 Short Form Citation

B13 Legislative Materials

B14 Administrative and Executive Materials

B15 Books and Other Nonperiodic Materials

- B15.1 Full Citation
- B15.2 Short Form Citation

B16 Periodical Materials

- B16.1 Full Citation
 - B16.1.1 Consecutively Paginated Journals
 - B16.1.2 Nonconsecutively Paginated Journals and Magazines
 - B16.1.3 Student-Written Work
 - B16.1.4 Newspaper Articles
- B16.2 Short Form Citation

B17 Court and Litigation Documents

- B17.1 Full Citation
 - B17.1.1 Abbreviation
 - B17.1.2 Pinpoint Citations
 - B17.1.3 Date
 - B17.1.4 Electronic Case Filings (ECF)
- B17.2 Short Form Citation

B18 The Internet

- B18.1 Full Citation
 - B18.1.1 Direct Citations
 - B18.1.2 Parallel Citations
- B18.2 Short Form Citation

B19 Services



B20 Foreign Materials

B21 International Materials

Challenges: sliced URLs

Outstanding warrants are surprisingly common. When a person with a traffic ticket misses a fine payment or court appearance, a court will issue a warrant. See, *e.g.*, Brennan Center for Justice, Criminal Justice Debt 23 (2010), online at <https://www.brennancenter.org/sites/default/files/legacy/Fees%20and%20Fines%20FINAL.pdf>. When a person on probation drinks alcohol or breaks curfew, a court will issue a warrant. See, *e.g.*, Human Rights Watch, Profiting from Probation 1, 51 (2014), online at <https://www.hrw.org/report/2014/02/05/profitting-probation/americas-offender-funded-probation-industry>. The States and Federal Government maintain databases with over 7.8 million outstanding warrants, the vast majority of which appear to be for minor offenses.

Challenges: link rot and content drift


- Nothing lasts forever on the web...
- Link rot in legal citations (Zittrain et al., 2014):
 - >70% of URLs in the Harvard Law Review = 
 - >50% of URLs in SCOTUS opinions = 
- Solution: perma.cc

CourtListener × Cobaltmetrics


- \gtrsim 3 million citations in US legal opinions
- Primary authority \approx 99%
 - Opinions, statutes, rules, regulations, and legislation
- Secondary authority \approx 1%
 - Everything else, e.g. dictionaries and short URLs

source-dataset:courtlistener domain:doi.org


26 citations

 **McKown v. Secretary of Health and Human Services**, United States Court of Federal Claims (via CourtListener), accessed August 29, 2019.

Matching URIs: [domain:doi.org](#)

 **J.H. and A.R. v. R & M Tagliareni, LLC (081128)(Hudson County and Statewide)**, Supreme Court of New Jersey (via CourtListener), accessed July 31, 2019.

Matching URIs: [domain:doi.org](#)

 **Petition of the United States on Behalf and for the Benefit of Smithsonian Institution**, District Court, District of Columbia (via CourtListener), accessed July 31, 2019.

Matching URIs: [domain:doi.org](#)

Y'all got any more of that?

<https://cobaltmetrics.com/docs/api>

A note on reproducibility

We aggregate data from different sources,
so there are **many moving parts**.

Our default strategy is to **ingest the entire datasets**.

Our API can return a **fingerprint** of the whole database,
as well as the **log of all the web resources** we remix.

Towards an open business model

- Currently mostly closed-source, but...
- **Open roadmap** on cobaltmetrics.com
- Everything on the website (data/docs) is **CC BY 4.0**



THUNKEN