

Measuring PID Adoption in Online Citations

Luc Boruta & Damien Vannson — Thunken
luc@thunken.com — @thunkenizer
Altmetrics19 Workshop, 2019/10/11



THUNKEN

Persistent identifiers

Nothing lasts forever on the web...

Link rot! Content drift! Outages!

PIDs were designed to slow down link rot.



<https://gph.is/2NehBG5>

Better a URL today than a PID tomorrow

The ideal identifier should be **persistent**,
findable, accessible, interoperable, and reusable...

...we all **copy-paste from the address bar** of our browser.

PIDs are not silver bullets

There are **billions of documents** that will never get DOIs or any other fancy PID: old documents, grey literature, and **the rest of the web**.

There are tons of documents with PIDs that are cited with no mention of their PIDs.

Why should we care?

Metrics are a sampling game.

Imbalanced datasets reinforce discrimination.

We are interested in **low-frequency phenomena**,
and in distinguishing **structural zeros** from **sampling zeros**.

Cobaltmetrics

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

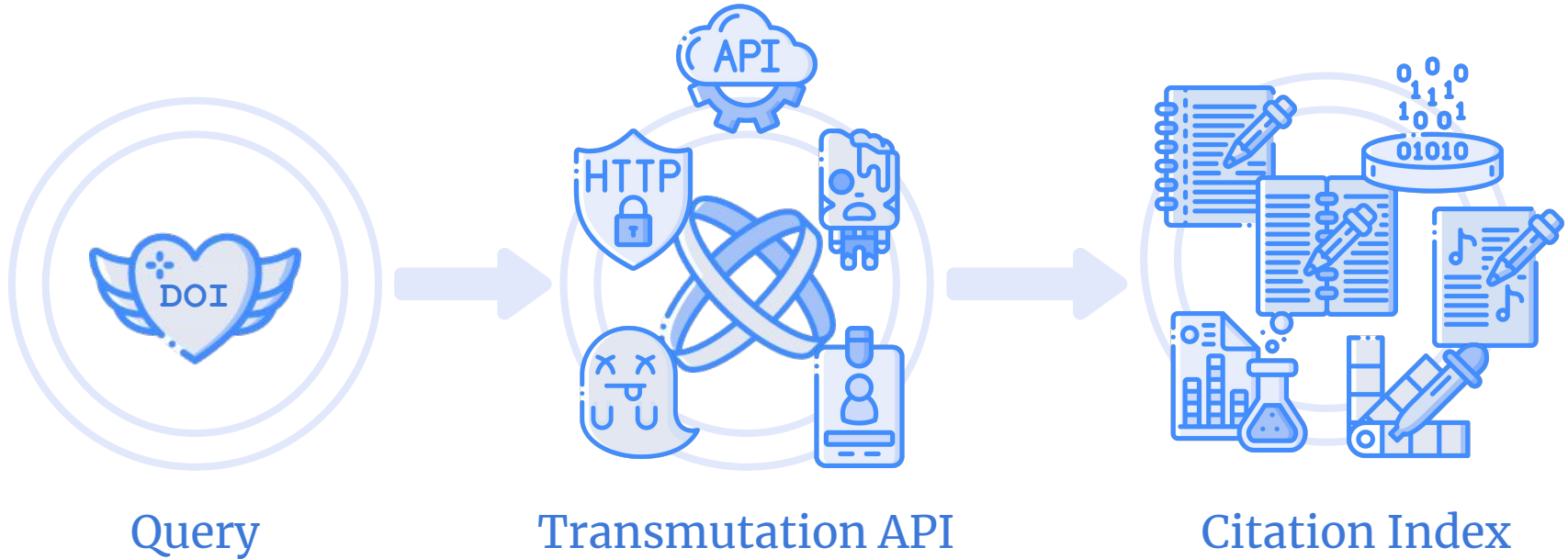


Query

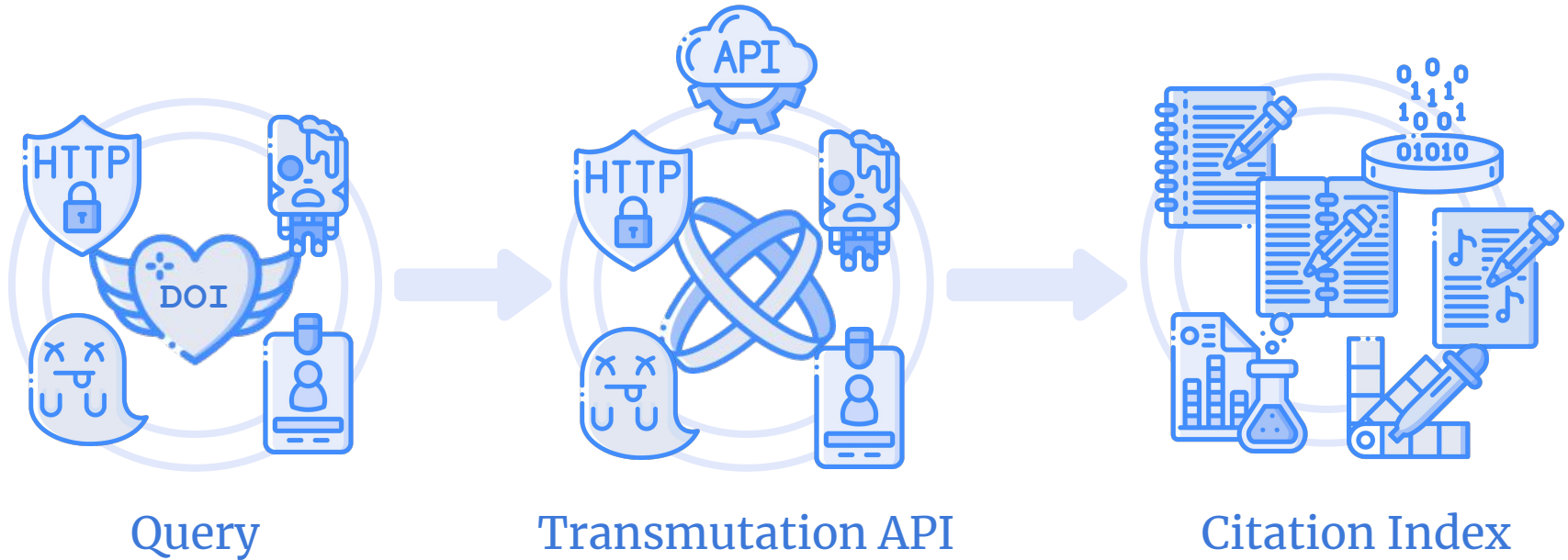


Citation Index

Cobaltmetrics



Cobaltmetrics



URI transmutation

Transmutation = normalization + conversion

- Equivalencies we can compute (e.g. ORCID $\rightleftharpoons$ ISNI)
- Equivalencies we must learn (e.g. short URL $\rightleftharpoons$ URL)

Our transmutation API is open and free, try it out!

Compact IDs vs. good old URLs

Cobaltmetrics' citation index (February 2019):

- HTTP+HTTPS+FTP: 256 million URLs (98%)
- Every other scheme: 4 million IDs

Non-canonical URIs

Non-canonical URI $\approx$ any ID that is not 100% FAIR,
including but not limited to:

- Short URLs
- Proxy URLs
- Sci-Hub URLs



URL taxonomy

1. Compact identifier
2. Resolver URL
3. Decorated resolver URL
4. Spliced URL
5. Extractable URL
6. URL



Compact identifier

- Scheme: anything but HTTP, HTTPS, Gopher, FTP
- *doi:10.5281/ZENODO.2548719*



Resolver URL

- Scheme: HTTP, HTTPS, FTP
- Non-empty host
- Path is a compact identifier
- Empty query, empty fragment
- ***<https://doi.org/10.5281/ZENODO.2548719>***



Decorated resolver URL

- Scheme: HTTP, HTTPS, FTP
- Non-empty host
- Path is a compact identifier
- *`https://doi.org/10.xxx/xxx?utm_source=newsletter`*



Spliced URL

- Scheme: HTTP, HTTPS, FTP
- Non-empty host
- Path/query/fragment contains a compact identifier
- *<https://aeaweb.org/articles.php?doi=10.1257/jep.30.1>*



Extractable URL

- Scheme: HTTP, HTTPS, FTP
- Our URI transmutation API returns at least one PID
- *<https://nature.com/articles/s41597-019-0026-5>*



URL

- ...any other hyperlink!

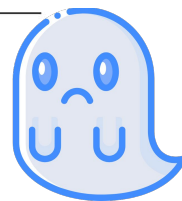
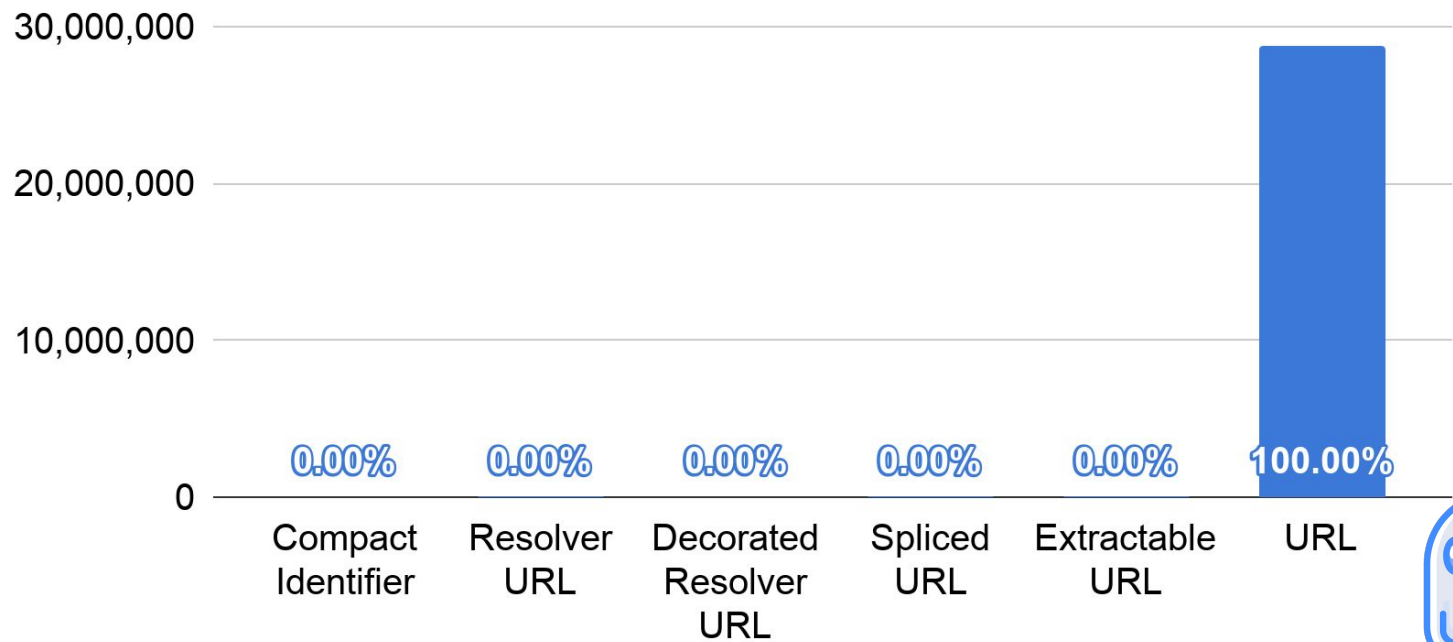


New API endpoint

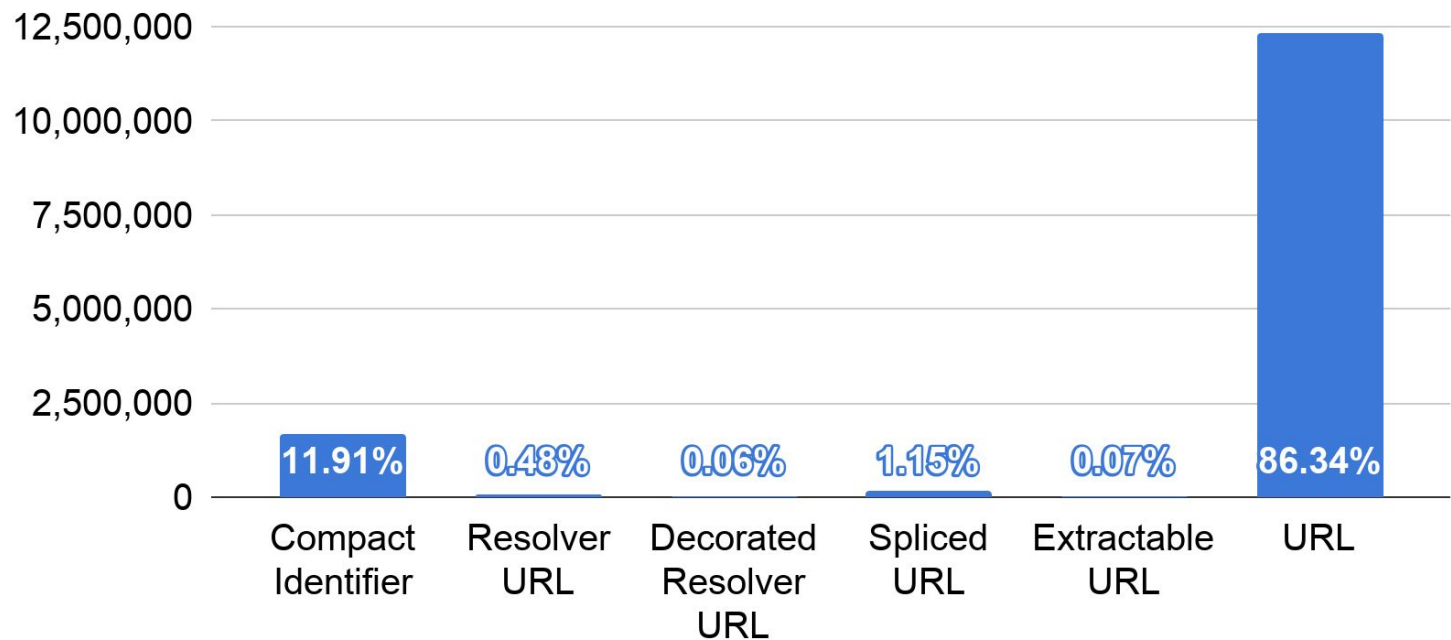
<https://cobaltmetrics.com/api/v2/citations/categories>



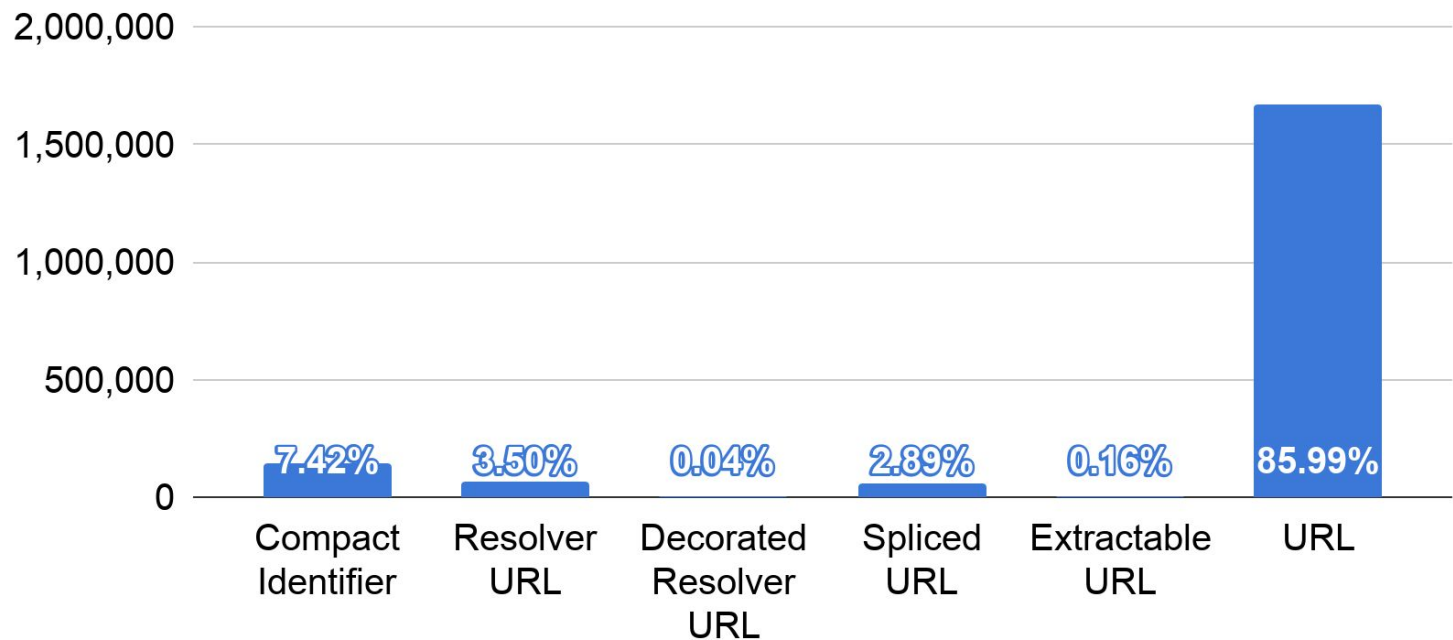
source-dataset:courtlistener



source-dataset:wikimedia



source-dataset:wikimedia language:eng



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.

Where do we go from here?

- Breakdown by domain/hostname
- More categories to refine the default “URL” category
- More resources (data/code) to improve URI transmutation
- Stats over time for a fixed dataset



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

Cobaltmetrics: web-scale citation tracking

Crossref	DOI, HTTP, HTTPS
GLEIF	BIC, LEI
GRID	GRID, ISNI, WIKIDATA
IETF	BCP, DOI, FYI, RFC, STD
ISSN	ISSN
NLM	ISSN, NLMID
ORCID	Email addresses, Google Scholar, ISNI, ORCID, ResearcherID, Scopus
PMC	DOI, PMCID, PMID
Ringgold	RIN, ISNI
Unpaywall	DOI, HTTP, HTTPS, OAI
URLTeam	FTP, HTTP, HTTPS
Wikidata	arXiv, Bibcode, CODEN, CVE, DOI, EIN, Google Books, Google Scholar, GRID, ISBN, ISNI, ISSN, JSTOR, LCCN, LEI, MR, NLMID, ORCID, PMCID, PMID, ResearcherID, RFC, RIN, SSRN, VIAF, YouTube, Zenodo
Wikimedia	HTTP, HTTPS