

Cobaltmetrics

Better a URL Today Than a PID Tomorrow

Luc Boruta — Thunken Inc.
luc@thunken.com — @thunkenizer
FREYA Ambassadors Webinar, 2019/09/24

<http://gph.is/X18Wen>



THUNKEN

A partial landscape of citation aggregators

- Journal to journal: Web of Science, Scopus
- DOI to DOI: OpenCitations
- URL to DOI: ALM/Lagotto, Crossref Event data
- URL to URL: Altmetric, Plum, **Cobaltmetrics**

Common issues with citation aggregators

- Imbalanced datasets
 - Predefined lists of supported research outputs
 - Predefined lists of supported languages
 - **Predefined lists of supported PIDs/URLs**
- Irreproducible indicators
 - Dependency on 3rd party servers (short URLs, APIs)

Why should we care?

Metrics are a sampling game.

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

The web is not FAIR (and will most likely never be) and **that is just fine.**

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: design rationale

Cobaltmetrics tracks all URIs, URLs, and typed PIDs.

Cobaltmetrics can only be queried by URIs.

Cobaltmetrics will never create new identifiers.

Cobaltmetrics will never create new metrics.

Cobaltmetrics: design rationale

- ✓ <http://dx.doi.org/10.1109/2.901164>
- ✓ [doi:10.1109/2.901164](https://doi.org/10.1109/2.901164)
- ✓ <https://ieeexplore.ieee.org/document/901164/>
- ✓ <https://bit.ly/2kEavO1>
- ✗ Lawrence et al., 2001

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.

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

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!

URI transmutation example

We remix 4M cliques of IDs from ORCID's Public Data File.

Example:

- `orcid:0000-0003-0557-1155` → `{scopus:55148973700}`
- `scopus:55148973700` → `{orcid:0000-0003-0557-1155}`
- `mailto:luc@thunken.com` → `{orcid:0000-0003-0557-1155, scopus:55148973700}`

Web-scale citation tracking

- Wikimedia (all projects, all languages)
- StackExchange/StackOverflow (all projects, all languages)
- US legal opinions (via CourtListener)
- Hypothes.is annotations
- Usenet posts (via the Internet Archive)
- **CommonCrawl (3.1 billion webpages)**

Web-scale citation tracking: transmutation

- **Crossref**
- ORCID
- PMC
- **Terror of Tiny Town**
- Unpaywall
- Wikidata
- ...

A note on reproducibility

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

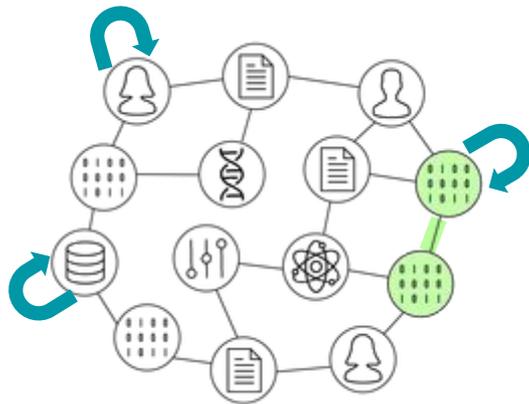
Our default strategy is to **ingest the entire datasets**, so that we control when and how data is updated.

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

Cobaltmetrics × FREYA

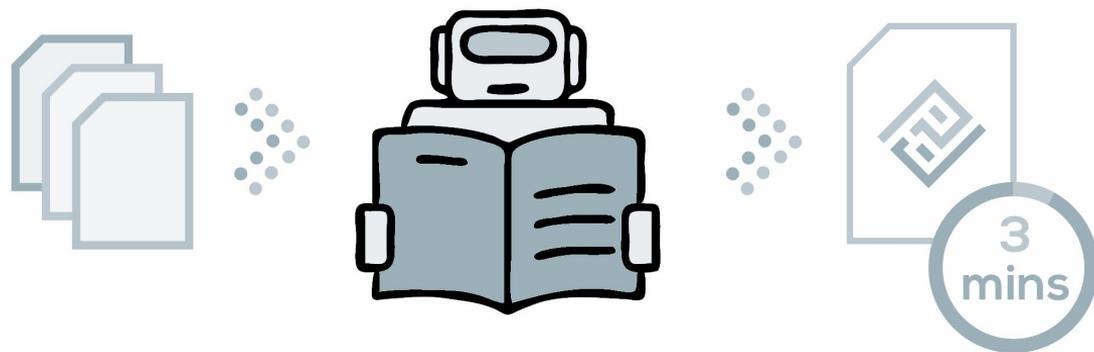
- **FREYA: PID graph**
 - Focus on scholarly entities
 - Many relations between entities
- **Cobaltmetrics: URI graph**
 - Scholarly entities, but not only
 - Only one relation between entities \approx owl:sameAs

Cobaltmetrics × FREYA



Cobaltmetrics × **FREYA**

Cobaltmetrics × Paper Digest



- AI-powered summaries
- Uses our transmutation API for inputs other than DOIs

Towards an open business model

- Currently mostly closed-source, but...
- Everything on the website (data/docs) is now **CC BY 4.0**
- Coming soon:
 - No more third party trackers
 - Pricing transparency



THUNKEN

Bonus: PIDs for RIs!

- MERIL-2 (EU Horizon 2020 grant, ESF/EKT/APRE)
 - DB of EU research infrastructures: <http://meril.eu/>
- MERIL-3 (RDA EU adoption grant, ESF/Thunken)
 - Customized DSpace + Handles
 - Blog post: <https://t.co/gNxgkDam7v>



THUNKEN