Cobaltmetrics
Web-Scale Citation Tracking

Luc Boruta & Damien Vannson — Thunken Inc.
luc@thunken.com — @thunkenizer
PUBMET2019, Zadar, 2019/09/20
Dear Santa,

How are you? I’m good.

Here is what I want for Christmas.

http://www.amazon.com/gp/product/B0032HF60

http://theinclusive.net/article.php?id=268
Attention vs. Impact

Citations and altmetrics are proxies for impact.

Citations and altmetrics measure attention.

Attention correlates w/ impact. So do influence and privilege.

Mentions and events are merely newish types of citations.
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
- Irreproducible indicators
  - Dependency on 3rd party servers (short URLs, APIs)
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.
Weapons of math destruction

“There is a moral obligation to challenge machine biases.”
— Heather Staines, PIDapalooza’19

Algorithmic bias reflects the values of the humans involved in designing the algorithm and/or collecting the data.
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 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.
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.
Design rationale

Lawrence et al., 2001, https://doi.org/10.1109/2.901164
http://dx.doi.org/10.1109/2.901164
doi:10.1109/2.901164
https://bit.ly/2kEavO1
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.
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
Are your metrics alt- enough?

NO.
Are your metrics alt-enough?

- Bias in favor of English
- Bias in favor of traditional publication venues
- Bias in favor of traditional publication formats
- Bias in favor of short-term rewards (vs. long-term goals)
- …?
Selection biases: Wikipedia languages

Altmetric: 3 languages (en, fi, sv)
PlumX Metrics: 3 languages (en, es, pt)
ALM: 25 most popular languages
Cobaltmetrics: 180+ languages!
Selection biases: document types

Strong focus on traditional peer-reviewed publications. Preprints are still treated as second-class documents.

What about patents, clinical trials, law articles, etc.? What about non-textual objects, e.g. datasets or software?

In Cobaltmetrics a URL is a URL, we do not discriminate.
Selection biases: PIDs vs. URLs

Nothing lasts forever on the web:

- Link rot!
- Content drift!
- Outages!
Non-canonical URIs

Non-canonical URI ≈ 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 ⇄ ISNI)
- Equivalencies we must learn (e.g. short URL ⇄ 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}
A note on reproducibility

Because we aggregate data from different sources, there are many moving parts.

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

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

https://cobaltmetrics.com/docs/page/data-sources
Cobaltmetrics in the context of open science

- 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