Cobaltmetrics

Web-Scale Altmetrics

Luc Boruta — Thunken Inc.
luc@thunken.com — @thunkenizer
GFII Open Science, 2019/03/13


THUNKEN

web-scale
alt-
-metrics

Are your metrics alt- enough?

NO.

Diversity in altmetrics

Altmetric, Event Data, and Plum are great projects.

But **diversity is good**, and we think we can do even better.

Are your metrics alt- enough?

- Bias in favor of **English**
- Bias in favor of **traditional publication venues**
- Bias in favor of **short-term rewards** (vs. long-term goals)
- …?

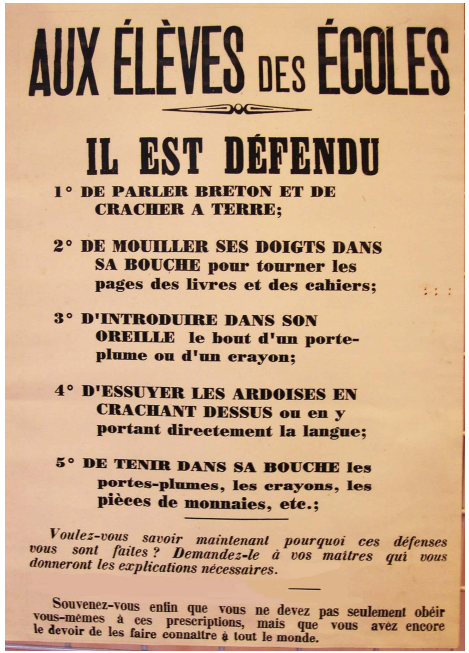
Toward fair and inclusive metrics

The scientific community is global and diverse.

Our corpus needs to be global and diverse:
all languages, all document types, all identifiers, etc.

It is not up to metrics providers to decide what is citable.

Why should we care?



AUX ÉLÈVES DES ÉCOLES

IL EST DÉFENDU

1° DE PARLER BRETON ET DE CRACHER A TERRE;

2° DE MOUILLER SES DOIGTS DANS SA BOUCHE pour tourner les pages des livres et des cahiers;

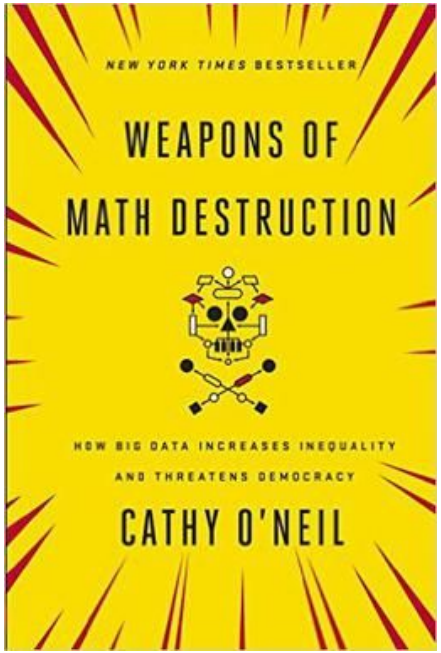
3° D'INTRODUIRE DANS SON OREILLE le bout d'un porte-plume ou d'un crayon;

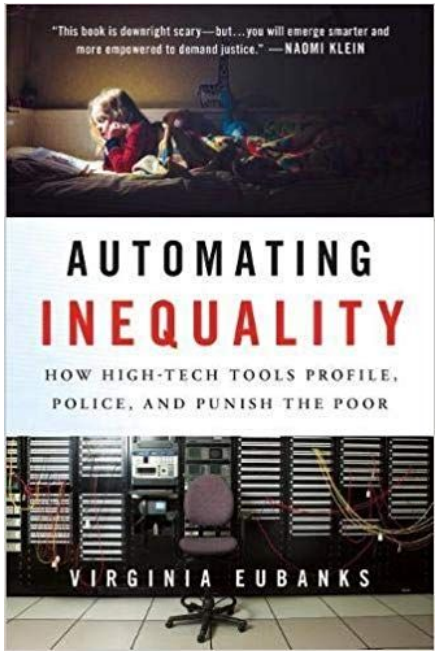
4° D'ESSUYER LES ARDOISES EN CRACHANT DESSUS ou en y portant directement la langue;

5° DE TENIR DANS SA BOUCHE les portes-plumes, les crayons, les pièces de monnaies, etc.;

Voulez-vous savoir maintenant pourquoi ces défenses vous sont faites? Demandez-le à vos maitres qui vous donneront les explications nécessaires.

Souvenez-vous enfin que vous ne devez pas seulement obéir vous-mêmes à ces prescriptions, mais que vous avez encore le devoir de les faire connaître à tout le monde.





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**.

Latent discrimination, real consequences

Systematic bias in stats on cross-linguistic citation practices.

Systematic exclusion of some contributors from metrics-based evaluations.

Reinforces discrimination in other parts of the community, cf. Nylenna et al. (1994) and Lazarev & Nazarovets (2018).

Selection biases in altmetrics: wiki languages

Altmetric: 3 languages (en, fi, sv)

PlumX Metrics: 3 languages (en, es, pt)

ALM: 25 most popular languages

Cobaltmetrics: 180+ languages!

https://medium.com/thunken/altmetrics-for-all-4b9039249850

Natural language processing to the rescue!

Anglo-centrism is prejudicial to science.

Algorithmic complexity cannot be used as an excuse.

Citation data is mostly **machine-readable**,
and/or can be described using **local grammars**.

Selection biases in altmetrics: 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?

https://medium.com/thunken/altmetrics-for-all-4b9039249850

DOIs 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 DOIs that are cited
with no mention of their DOI.

Web-scale altmetrics

- Wikimedia (all projects, all languages)
- StackExchange/StackOverflow
- Hypothes.is annotations
- Usenet posts (via the Internet Archive)
- Legal opinions (via CourtListener)
- **CommonCrawl (3+ billion webpages)**

https://cobaltmetrics.com/docs/page/data-sources

Cobaltmetrics

cobaltmetrics.com #altmetricsforall

luc@thunken.com — @thunkenizer