Crossref Event Data
A New, Open
Underlying Data Set
for Altmetrics

altmetrics17 workshop
Hello again!
What is an “underlying dataset”? 
Problems to tackle when building a metric or doing research.
Identifying a data set.
Identifying a data set.

Links from where?
Identifying a data set.

Links to where?
Identifying a data set.

What kind of links?
Non-traditional publications, non-traditional rules.
Collecting data from the corpus.
Identifying & canonicalising links.

Link rot in your dataset.
Recording data for reproducibility.

“The data we got are not transparent”
Who’s human?
Who’s a machine?
Two categories.
Category A

Everyone has to solve these.
Category B

Everyone has to different issues.
Category B

Particular corpus of articles. Links from a particular website. Blogs from known authors. Links via DOIs or not. Building different things?
Category A: We can help you solve.

Category B: We can do.
Crossref
~100 million items. Metadata. Infrastructure to make it work.
Metadata typically provided by publishers.
Citations

relations, funding information, author links

Crossref Registered Content

The Rest of the world
Crossref Registered Content

Citations

The Rest of the world

Relations, funding information, author links

Event Data
No longer supplied by publishers.
An underlying data set.

(Category A)
Problems
Which data set to use?
Which link source corpus?
• Reddit ~ 100,000
• Twitter ~ 3,000,000
• Wikipedia ~ 30,000,000
• hypothes.is ~ 20,000
• wordpress.com ~ 30,000
• Blogs ~ 200,000
• Stack Exchange ~ 6,000
What kind of links?
Canonicalisation and link rot.
Reproducibility of data collection.
Reproducibility of state.
Reproducibility of data processing.
Representing an evolving set of links.
Telling humans from machines.
Agents monitor external services

Percolator consults external services

Agents consume Artifacts

Agents produce Evidence Records

Percolator consumes Evidence Records

Percolator produces Events

Evidence Logs
All data is free. CC0, CC-BY-4.0, etc
Open for Beta
https://www.eventdata.crossref.org/guide/

https://www.crossref.org/blog

4:AM Talk & Hackathon!

eventdata@crossref.org

@joewass