Cobaltmetrics

Preventing citation decay and obfuscation

Luc Boruta, Thunken Inc.
luc@thunken.com — @thunkenizer
Workshop on Open Citations, Bologna, 2018/09/03

Preventing citation decay and obfuscation

1. Cobaltmetrics and URI transmutation
2. Proxy/short URLs: how many citations are we missing?
3. Dead links are not necessarily dead ends

Cobaltmetrics: design rationale

Cobaltmetrics tracks all URIs.

Cobaltmetrics can only be queried by URIs.

Cobaltmetrics will never create new identifiers.

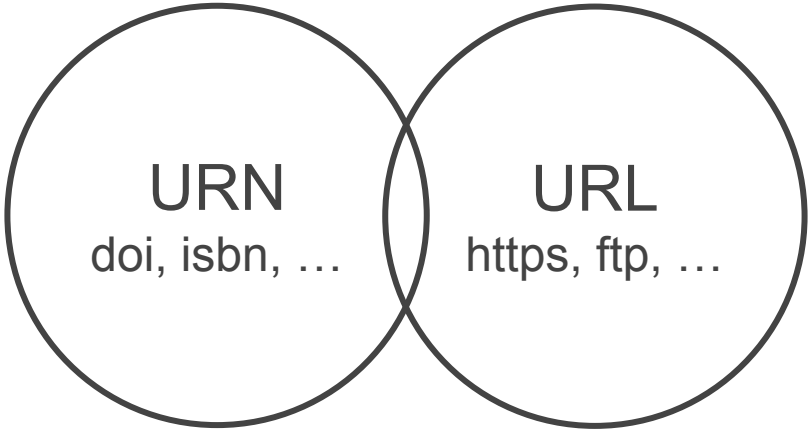
Cobaltmetrics will never create new scores.

Standardized IDs are useful, but not sufficient

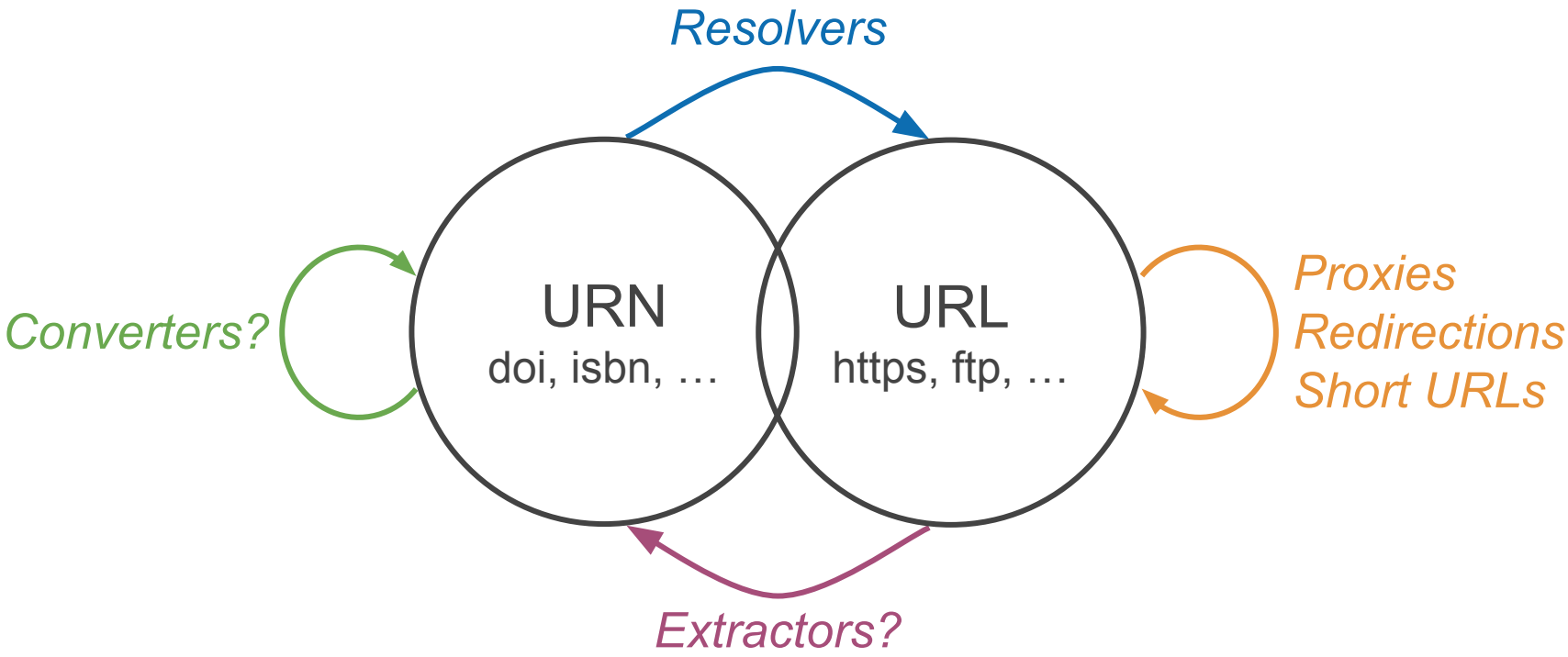
The ideal identifier should be **persistent**,
findable, accessible, interoperable, and reusable...

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

URI transmutation



URI transmutation



URL shortening was an awful idea

http://**goldbook**.iupac.org/html/A/**A00046**.html

https://doi.org/**10.1351/goldbook.A00046**

http://bit.ly/**2DZINQT**

http://doi.org/**ftz65g**

How many citations are we missing?

Which domains to track?

Can we use CUFTS? EZProxy? Wikidata?

First results using domains listed in EZProxy stanzas and Wikipedia, and a sample of 1.6B short URLs.

How many citations are we missing?

Medline/PubMed: 88k

Springer-Nature: 64k

Elsevier: 61k

Wiley-Blackwell: 50k

SAGE: 40k

DOI.org: 18k

Taylor & Francis: 18k

Sci-Hub is down, long live Sci-Hub!

Sci-Hub URLs are now used in published papers.

https://torrentfreak.com/sci-hub-proves-that-piracy-can-be-dangerously-useful-180804/

Sci-Hub domains are often suspended, creating dead links.

We support URLs that use any of Sci-Hub's domains,
active and deactivated domains alike.

Preventing citation decay

Supporting **deactivated domain names** is crucial.

Compare it to studying extinct languages:
even if they no longer have any users,
we still want to understand existing records.

Nothing lasts forever on the web*, so what can we do?

*except for that embarrassing picture on your old MySpace profile

Preventing citation decay

Secondary content providers should not use custom IDs
unless they release open mappings.

https://journals.**plos.org**/plosmedicine/article?id=**10.1371/journal.pmed.0020124**
https://www.**ncbi.nlm.nih.gov**/pmc/articles/**PMC1182327**/
https://**f1000.com**/prime/**8225**
https://www.**academia.edu**/**26252588**/Why_most_published_research...
https://www.**researchgate.net**/publication/**7686290**_Why_Most_Published_Research...
https://www.**scienceopen.com**/document?vid=**0d093096-2cbc-4feb-ba0e-07bf573eac5e**

Dead links are not necessarily dead ends

Academic publishers and libraries have **CLOCKSS**,
but we don't need to duplicate the metadata.

URL shorteners have **301Works.org**, the **URLTeam**,
and the **BEACON** format for large numbers of uniform links.

Secondary content providers should release BEACONs.

Lots of BEACONs keep stuff safe

```
#FORMAT: BEACON
#PREFIX: https://www.scienceopen.com/document?vid=
#TARGET: https://doi.org/
c8a48901-d9fd-49de-9378-7c64a025256b|10.1073/pnas.1320040111
0d093096-2cbc-4feb-ba0e-07bf573eac5e|10.1371/journal.pmed.0020124
34612198-44a2-4135-bb3b-cf4f5d8f93db|10.1038/srep01742
```


Dead links are not necessarily dead ends

Secondary content providers should release BEACONs,
or we can crawl the web and build them ourselves.

Open-source **parser**: https://github.com/thunken/beacon

Coming soon: Scholar-like **crawler** that outputs BEACONs.

Cobaltmetrics

cobaltmetrics.com #altmetricsforall

Let's build BEACONs together!

luc@thunken.com — @thunkenizer