Measuring Persistent Identifier Adoption in Online Citations
Luc Boruta\(^a\) & Damien Vannson\(^b\), Thunken Inc., Washington DC, USA
\(^a\)luc@thunken.com, https://orcid.org/0000-0003-0557-1155
\(^b\)damien@thunken.com, https://orcid.org/0000-0001-8387-7358

Objectives
Persistent identifiers (henceforth PIDs) are designed to slow down—and, ideally, stop—link rot for both scientific and non-scientific digital data. In the scientific community, guiding principles for scientific data management like the FAIR Principles\(^2\) or the Metadata 2020 Principles\(^3\) directly or indirectly advocate the use of PIDs such as DOIs for published works, ISSNs for journals, or ORCIDs for contributors. That being said, PIDs are not silver bullets. There are billions of documents and digital objects that will never be assigned a PID—e.g. works published before the advent of DOIs and most of the works that fall under the "grey literature" label—and documents that were assigned PIDs are not necessarily cited with mentions of these PIDs. The latter point is the most important in the context of altmetrics. Metrics are a sampling game: selection biases are an issue, and imbalanced datasets reinforce discrimination. Which citation patterns must we monitor to produce a corpus that is both diverse and inclusive? Are PID citations enough? Should web pages (e.g. landing pages on publisher websites) be included? What about non-canonical identifiers like proxy URLs (from services like EZproxy or Sci-Hub) or short URLs?

Traditional citation indexes have stringent inclusion criteria and focus on privileged (i.e. any combination of peer-reviewed, English-language, and/or PID-ready) publication venues. Altmetrics were designed to overcome some of these limitations, but most data providers still somehow rely on predefined lists of citable/indexable research outputs\(^4\) or acceptable citation patterns\(^5\). With Cobaltmetrics\(^6\), Thunken is on a mission to make altmetrics genuinely alternative. One of our core principles is that it is not up to altmetrics data providers to decide what is citable, our role is to observe most if not all citation patterns on the web. The web is not FAIR—and will most likely never be—and that is just fine.

In the hypertext era, creating actionable citations requires minimal effort. We argue that any hyperlink is a valid citation, and that tracking and disambiguating citations can be—and already is—outsourced to machines. Moreover, persistent identifiers and long-term archiving strategies can give an illusion of permanence, but most citations on the web are fragile\(^7\). We argue that the only way forward is to embrace web-scale citation tracking. To produce a corpus that is diverse and inclusive, we track all URIs: every hyperlink, every occurrence of a URI is collected as a citation in Cobaltmetrics. One of our biggest challenges...

---

1 Conflict of interest statement: Cobaltmetrics is a project of Thunken Inc., of which the authors are employees.
6 https://cobaltmetrics.com/
is to collate URIs that directly or indirectly identify the same resource, so that citation counts and attention scores can be tallied accurately.

Methods

To measure PID adoption in online citations, and to evaluate how much effort altmetrics aggregators should put into monitoring non-canonical citations—i.e. citations that use non-canonical or non-persistent identifiers—we collected statistics from the Cobaltmetrics corpus. The latest release of the corpus consists of citation data extracted from the following sources: CommonCrawl, CourtListener, Hypothesis, StackExchange, Usenet, and Wikimedia.8

Then, in order to collate URIs that directly or indirectly identify the same resource, we rely on a process called URI transmutation.9 URI transmutation can be defined as the combination of three operations on URIs:

- **URI normalization** (a.k.a. canonicalization or standardization), which is defined in part in the RFC for the generic syntax of URIs10, and further refined in other documents, depending on the scheme of the input URI;
- **URI equivalence**, which relies on linked open data resources and ontology mappings with properties such as owl:sameAs;
- **URI interpolation**, which are additional normalization operations that make simplifying assumptions on URIs and URI equivalence. We call them interpolations because they allow us to simplify the transmutation process and increase the number of matches, while not affecting any of the important conclusions drawn from the data.

URI transmutation can be thought of as a graph of URIs that is overlaid onto our citation index. In other words, for each citation, our citation index contains the exact URI that was used in the citing document, and the URI transmutation process allows us to access the citation using any URI known to be equivalent to the citing URI.

Citations in the corpus are classified according to the following typology of mutually-exclusive categories, listed here from most canonical to least canonical:

1. **Compact identifier**: a URI whose scheme is not http, https, gopher, or ftp, e.g. doi:10.5281/ZENODO.2548719;
2. **Resolver URL**: a URI whose scheme is http, https, or ftp, with a non-empty host component, a path component which is itself a compact identifier (with or without the identifier scheme), an empty query component, and an empty fragment component, e.g. https://doi.org/10.5281/ZENODO.2548719;
3. **Decorated resolver URL**: a URI whose scheme is http, https, or ftp, with a non-empty host component and a path component which is itself a compact identifier (with or without the identifier scheme), e.g. https://doi.org/10.5281/ZENODO.2548719?utm_source=newsletter;
4. **Spliced URL**: a URI whose scheme is http, https, or ftp, with a non-empty host component, and whose path component, query component, or fragment component contains a compact identifier (with or without the identifier scheme), e.g. https://www.aeaweb.org/articles.php?doi=10.1257/jep.30.1;

---

5. Extractable URL: a URI whose scheme is http, https, or ftp, and for which our URI transmutation API returns at least one equivalent URI, e.g. https://www.nature.com/articles/41597-019-0026-5;

6. URL: any URI that does not fall into any of the other categories.

Finally, in order to reduce the dimensionality of the data, the results are aggregated by data source and by publisher, using Crossref member identifiers to link works to publishers.

Results

Web-scale citation tracking is a new and daunting endeavor and—full disclosure—we do not have results at the time of writing! Ha! This is the first time we dare submit a work in progress, but we hope the approach is novel enough to appeal to your curiosity.

We have set up the infrastructure and the code is ready to run, but we just signed the metadata agreement with Crossref today (July 22, 2019) to expand our knowledge graph, run a full-scale analysis, and aggregate the results. We will share the lessons we have learned in the past 18 months, including negative results and tips to pull free citation data with our API. We will conclude with a special announcement about our APIs, linked open data, and permissive data licenses!