Unlocking citations from tens of millions of scholarly papers

Dario Taraborelli

SWIB 2017 • Hamburg, 6 December 2017
Did my lecturer just cite a source as Wikipedia? Wild.
Aaron Chatman
@AaronChatman1

I'll donate to Wikipedia when I can apa cite them

1:27 PM - 4 Dec 2017

13 Likes
Wikipedia: Verifiability, not truth

From Wikipedia, the free encyclopedia

This essay contains the advice or opinions of one or more Wikipedia contributors on the Verifiability policies. Essays are not Wikipedia policies or guidelines. Some essays represent widespread norms; others only represent minority viewpoints.

This page in a nutshell: Any material added to Wikipedia must have been published previously by a reliable source. Editors may not add content solely because they believe it is true, nor delete content they believe to be untrue, unless they have verified beforehand with a reliable source.

Wikipedia's core sourcing policy, Wikipedia: Verifiability, used to define the threshold for inclusion in Wikipedia as "verifiability, not truth". "Verifiability" was used in this context to mean that material added to Wikipedia must have been published previously by a reliable source. Editors may not add their own views to articles simply because they believe them to be correct, and may not remove sources' views from articles simply because they disagree with them.

en.wikipedia.org/wiki/Wikipedia:Verifiability,_not_truth
Published in 1609, the Sonnets were the last of Shakespeare's non-dramatic works to be printed, but evidence suggests that Shakespeare wrote sonnets
[132] Even before the two unauthorised sonnets appeared in The Passionate Pilgrim in 1592, Shakespeare's "sugred Sonnets among his private friends".[133] Few analysts believe that they formed an intended sequence.[134] He seems to have planned two contrasting series: one about unrequited love and another about the inevitable passing of time. The "fair youth" he refers to in the sonnets may represent real individuals, or if the authorial "I" who addresses them represents Shakespeare himself, it was in jest that he composed the sonnets "Shakespeare unlocked his heart".[135]

The 1609 edition was dedicated to a "Mr. W.H.", credited as "the only begetter" of the poems. It is not known whether this was written by Shakespeare himself or by the publisher, Thomas Thorpe, whose initials appear at the foot of the dedication page; nor is it known who Mr. W.H. was, despite numerous theories, or whether Shakespeare even authorised the publication of the sonnets.
[136] Critics praise the Sonnets as a profound meditation on the nature of love, sexual passion, procreation, death, and time.[137]
provenance
funding
Scopus®

Scopus is the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings.

Web of Science is the most comprehensive resource – we value both quality & quantity.

We are independent and unbiased.
"It is a scandal that mass access to citation data is still in the hands of a small group of closed-access players". –@dshotton

#WikiCite
The Initiative for Open Citations (I4OC)

Initiative for Open Citations

The Initiative for Open Citations I4OC is a collaboration between scholarly publishers, researchers, and other interested parties to promote the unrestricted availability of scholarly citation data.
The aim of this initiative is to promote the availability of data on citations that are \textit{structured}, \textit{separable}, and \textit{open}.

\textbf{Structured} means the data representing each publication and each citation instance are expressed in common, machine-readable formats, and that these data can be accessed programmatically. \textbf{Separable} means the citation instances can be accessed and analyzed without the need to access the source bibliographic products (such as journal articles and books) in which the citations are created. \textbf{Open} means the data are \textit{freely accessible and reusable}. 
How it came together
How it came together

**The starting point**
Most publishers already deposit their reference data with Crossref
The default state for the data is closed

**The challenge**
Could we persuade a group of influential publishers to release their data all at once?
Making the case

It’s easy and doesn’t cost anything
All you need to do is to send an email to support@crossref.org

The goal cannot be achieved alone
A comprehensive network of all scholarship can only be achieved if data is pooled

Publishers also benefit
Better discovery tools mean that content will be found and used more
Making it happen

Focus on publishers depositing the most data
Contacted the top-20 publishers asking for agreement in principle and permission to share their decision

Agree a deadline
Everyone has time to prepare their comms and to be part of a big splash

Leverage the early adopters
As soon as we had a few publishers on board, others quickly followed
Progress so far
How many citations are open today?

1%
How many citations are open today?
Progress

18 million

DOI records with open references
Progress

500 million open reference data points
STAKEHOLDERS OF THE INITIATIVE FOR OPEN CITATIONS • https://i4oc.org/#stakeholders
Data reuse

The Open Citations Corpus

A broad and open collection of citation information from many sources

David Shotton and Silvio Peroni

THE OPEN CITATIONS CORPUS • http://opencitations.net/corpus
Data reuse

VISUALIZING FREELY AVAILABLE CITATION DATA USING VOSVIEWER • https://www.cwts.nl/blog?article=n-r2r294
Data reuse

The Wikidata Citation Graph

36 million citation links using the cites (P2860) Property in Wikidata

PARTIAL CITATION GRAPH FOR ULRICH K. LAEMMLI (1970) · http://tinyurl.com/y7acpqzd
Data reuse

*Tools to create profiles*

Scholia uses data from Wikidata

PROFILE INFORMATION FOR **EGON WILLIGHAGEN** •  [https://tools.wmflabs.org/scholia/author/Q20895241](https://tools.wmflabs.org/scholia/author/Q20895241)
The road ahead
Lessons learned

A single, measurable goal
Low cost
Agnostic to business model
Amplification
Towards an open graph for scholarship

“The visualization shows a structure of science that is well known from earlier large-scale bibliometric visualizations, which were based on Web of Science or Scopus data.”

VISUALIZING FREELY AVAILABLE CITATION DATA USING VOSVIEWER • https://www.cwts.nl/blog?article=n-r2r294
Who benefits from this

Q. What do you see as being the key benefits for authors and researchers of a fully open citation dataset?
A. The availability of this data benefits authors, researchers, funding and evaluation bodies, publishers, and the general public alike.

- **Authors** will have consistent, machine-readable access to references for all their publications;
- **Researchers** will be able to use this resource to study the dissemination of methods and scientific ideas, the genesis and provenance of scholarly knowledge;
- **Funders** will be able to rely on a public resource to develop transparent and reproducible evaluation metrics, and new tools to assess the academic and societal impact of research they fund;
- **Publishers** will benefit from the increased discoverability of publications that this data provides, and tools built on it.
- **The public** will be able to use this data to trace knowledge back to its sources or reuse it in open knowledge repositories such as Wikipedia and Wikidata.
Challenges: coverage

41% Crossref records have reference data

47% of those have open reference data

Challenges: data quality

Over 1 billion references
49% are open
53% have DOIs (and can be linked to another record)

The road to 100%
The road to 100%

Major publishers among the top 20 DOI depositors *not distributing open references* (as of October 2017)

- Elsevier
- IEEE
- Wolters Kluwer Health
- IOP Publishing
- Oxford University Press
- American Chemical Society
# The road to 100%

A list of all Crossref members with open references and statistics on their open reference coverage

<table>
<thead>
<tr>
<th>Member Name &amp; ID</th>
<th>Sponsored member &amp; prefix</th>
<th>Open References</th>
<th>Total Backfile DOIs</th>
<th>Total Current DOIs</th>
<th>Deposits Backfile References</th>
<th>Deposits Current References</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIP Publishing (ID 317)</td>
<td>American Institute of Physics 10.1063</td>
<td>true</td>
<td>674618</td>
<td>49449</td>
<td>true</td>
<td>true</td>
</tr>
<tr>
<td>AOSIS (ID 2580)</td>
<td>AOSIS 10.4102</td>
<td>true</td>
<td>20226</td>
<td>2700</td>
<td>true</td>
<td>true</td>
</tr>
<tr>
<td>Alliance Against Traffic in Women Foundation (ID 5611)</td>
<td>Alliance Against Traffic in Women Foundation 10.14197</td>
<td>true</td>
<td>64</td>
<td>48</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>American Association for the Advancement of Science (AAAS) (ID 221)</td>
<td>American Association for the Advancement of Science (AAAS) 10.1126</td>
<td>true</td>
<td>264316</td>
<td>13047</td>
<td>true</td>
<td>true</td>
</tr>
</tbody>
</table>

CROSSREF MEMBERS WITH OPEN REFERENCES • [https://www.crossref.org/reports/members-with-open-references/](https://www.crossref.org/reports/members-with-open-references/)
“This is a matter of scientific integrity, scientific progress, and equity—we must ensure that all members of the scientometric community are able to participate in and validate the research in the field. I4OC is striving to create such an opportunity.

– Cassidy R. Sugimoto et al.
Getting involved

An Open Letter to Stakeholders of the #I4OC: Help us make all indexed scholarly citation data openly available
i4oc.org/news.html#Augu...

An Open Letter to Stakeholders of the Initiative for Open Citations

Dear I4OC Stakeholders,

It's now four months since we publicly announced the Initiative for Open Citations (I4OC). Since the beginning of this effort, almost half of indexed scholarly citation data have become freely accessible. We've also had some amazing initial press coverage and we continue to add new publishers and stakeholders.

Data unlocked by I4OC is already being used by a growing number of projects and platforms. OpenCitations imports citation data into a corpus which now includes more than 9 million citation links, a nearly 200% increase since the beginning of the year. Collaborative databases, such as Wikidata, are already using this data to connect and structure knowledge and to generate citation graphs. These examples provide just an early indication of the potential of open citation data and we would be delighted to hear about other efforts.

I4OC's progress so far has been achieved thanks to helpful conversations with many of the larger publishers, and the majority have already decided to make their references freely available. But there are literally hundreds more publishers who are not currently making their reference data available even though this data is deposited with Crossref. We suspect this is largely because these organisations don't realise that citation data is closed by default.

7:51 AM - 8 Aug 2017

https://twitter.com/i4oc_org/status/894934190625402880
Thank you

D. Taraborelli (2017) Unlocking citations from tens of millions of scholarly papers

*SWIB 2017* [CC BY 4.0](https://doi.org/10.6084/m9.figshare.5674486)

**Acknowledgments**

The *I4OC founders*: OpenCitations, Wikimedia Foundation, PLOS, eLife, DataCite, the Center for Culture and Technology at Curtin University.

The *I4OC instigators*: Jonathan Dugan, Martin Fenner, Jan Gerlach, Catriona MacCallum, Daniel Mietchen, Cameron Neylon, Mark Patterson, Michelle Paulson, Silvio Peroni, David Shotton. Daniel Ecer for data analysis of the Crossref corpus.

The *I4OC stakeholders* ([i4oc.org/#stakeholders](https://i4oc.org/#stakeholders)) and participating publishers ([i4oc.org/#publishers](https://i4oc.org/#publishers))