Unlocking citations from tens of millions of scholarly papers

Dario Taraborelli

SWIB 2017 •Hamburg, 6 December 2017

Mavis Smith

@SunshineCityKid

Did my lecturer just cite a source as Wikipedia? Wild.

5:47 AM - 5 Dec 2017

Aaron Chatman
@AaronChatman1

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

1:27 PM - 4 Dec 2017

13 Likes

1 13

Wikipedia:Verifiability, not truth

From Wikipedia, the free encyclopedia

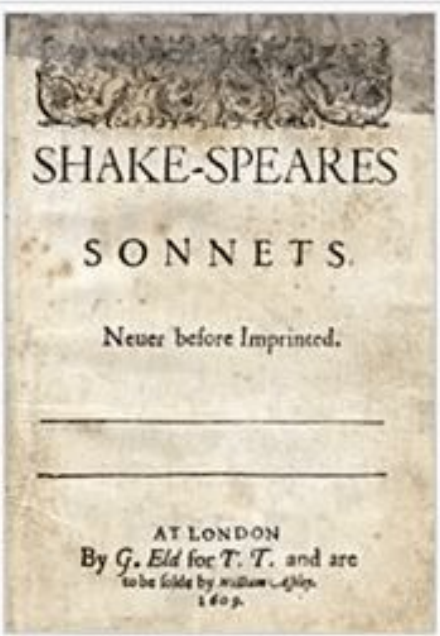
This essay contains the advice or opinions of one or more Wikipedia contributors on the Verifiability policies. Essays are *not* Wikipedia policies or guidelines. Some essays represent widespread norms; others only represent minority viewpoints.

Shortcuts:
WP:NOTTRUTH
WP:TRUTH
WP:VNT

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


SHAKE-SPEARES

SONNETS.

Neuer before Imprinted.

AT LONDON
By G. Eld for T. T. and are
to be solde by William Aspley.
1609.

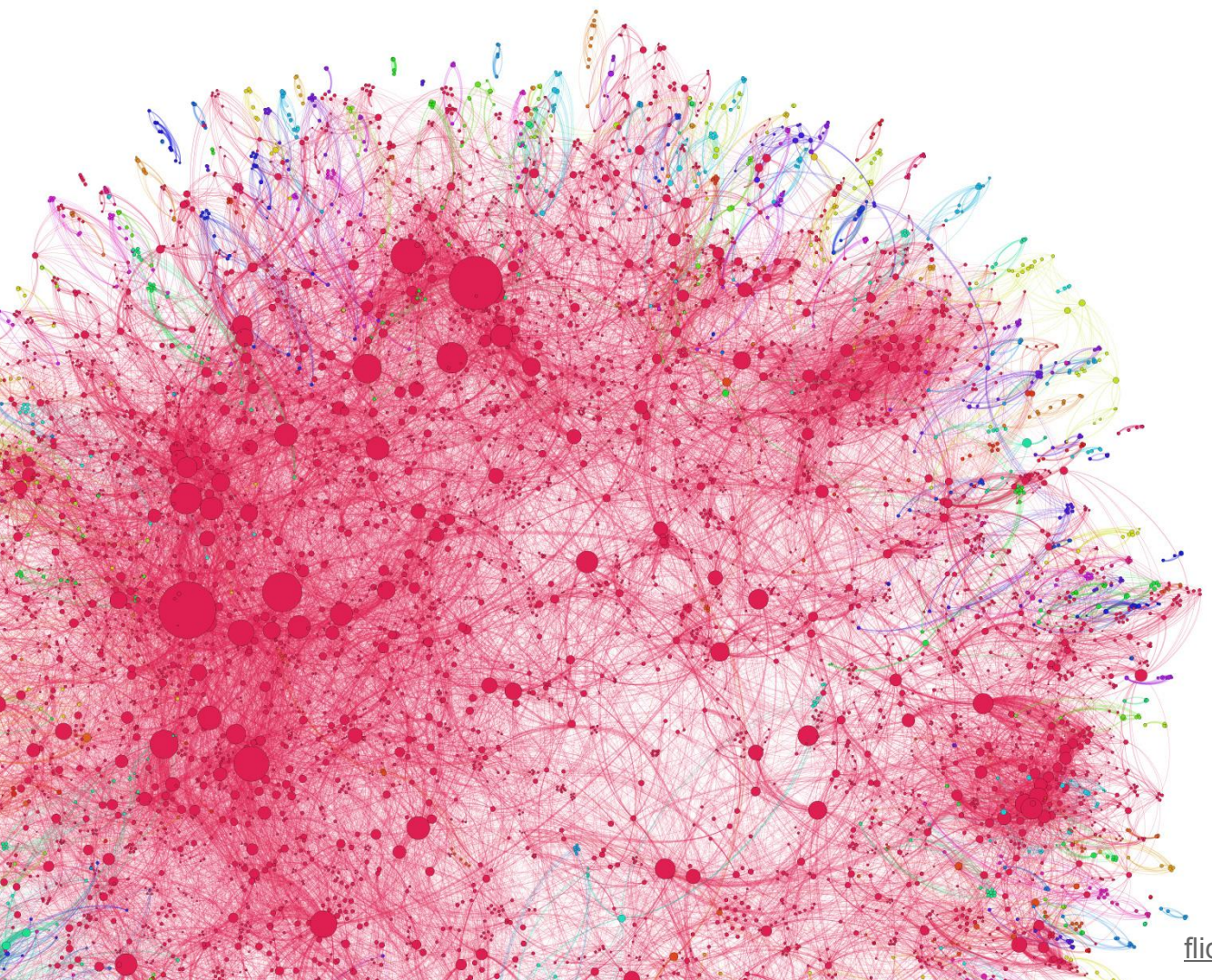
Title page from 1609 edition of *Shake-Speares Sonnets*.

Published in 1609, the *Sonnets* were the last of Shakespeare's non-dramatic works to be the 154 sonnets was composed, but evidence suggests that Shakespeare wrote sonnets [132] Even before the two unauthorised sonnets appeared in *The Passionate Pilgrim* in 15 Shakespeare's "sugred Sonnets among his private friends".[133] Few analysts believe tha intended sequence.[134] He seems to have planned two contrasting series: one about un complexion (the "dark lady"), and one about conflicted love for a fair young man (the "fair represent real individuals, or if the authorial "I" who addresses them represents Shakesp with the sonnets "Shakespeare unlocked his heart".[135]

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

Clear

Style



provenance


impact
SPACE X

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**.

WikiCite
@Wikicite

"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

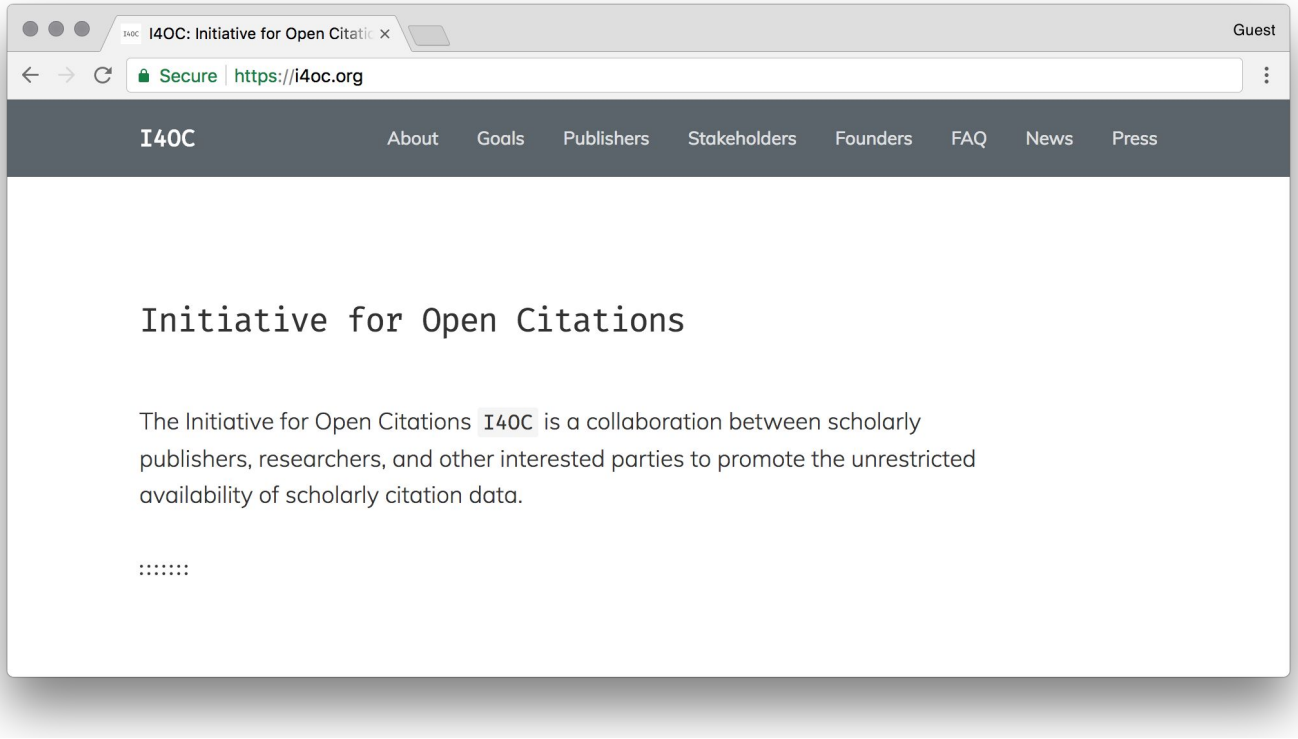
5:25 AM - 23 May 2017

24 Retweets **23** Likes

DEAN MORLEY [CC BY ND] • flickr.com/photos/33465428@N02/4490667565

BLUESTAR FRUIT RIPENING GAS • indiamart.com/cold-room-engineers/

The Initiative for Open Citations (I4OC)



The Initiative for Open Citations (I4OC)

The aim of this initiative is to promote the availability of data on citations that are **structured**, **separable**, and **open**.

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. **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. **Open** means the data are freely accessible and reusable.

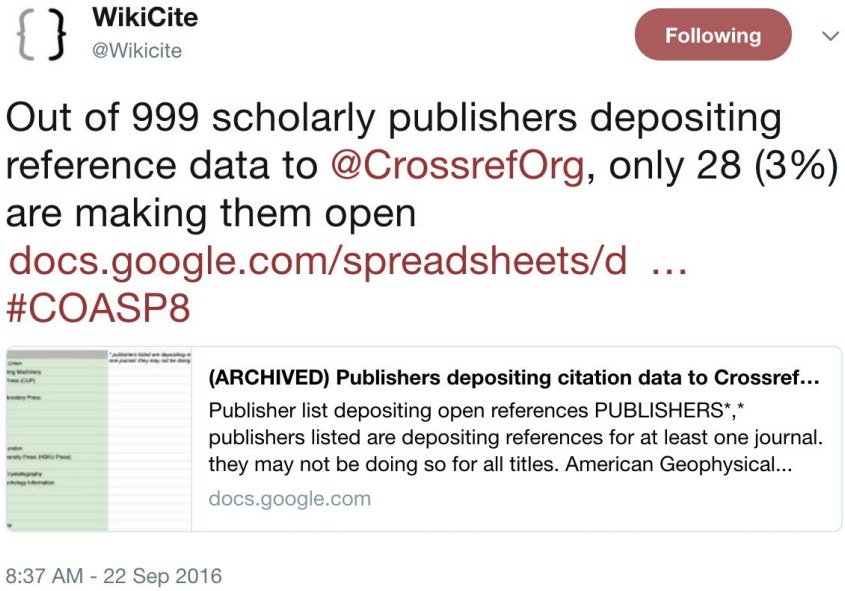
How it came together

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

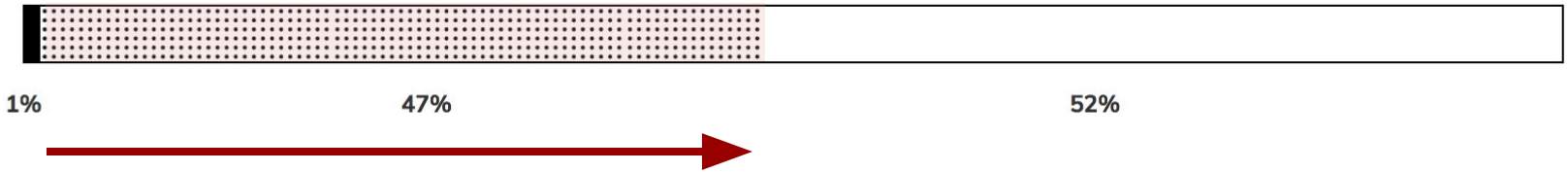
Progress

How many citations are open today?

1%

Progress

How many citations are open today?



Progress

18 million

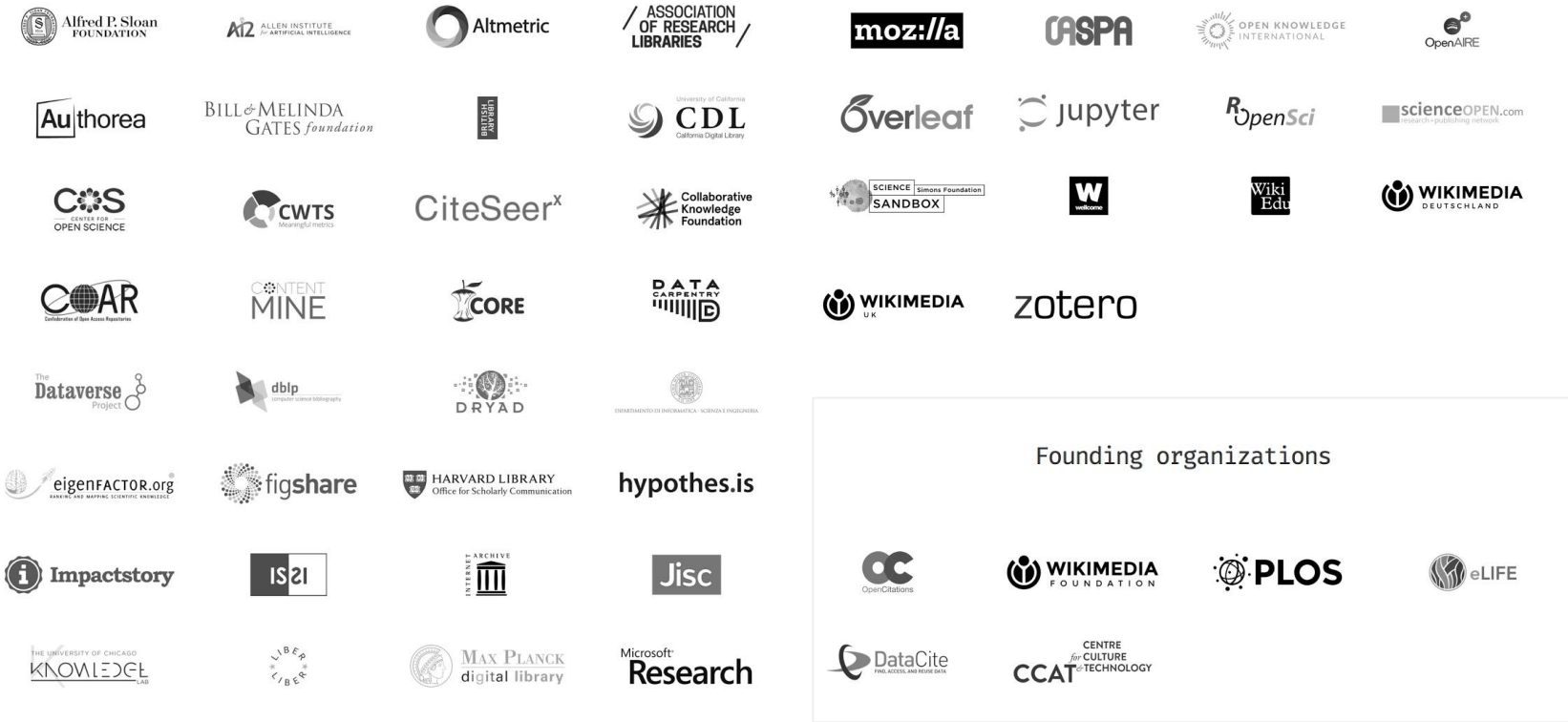
DOI records with open references

Progress

500 million

open reference data points

Stakeholders



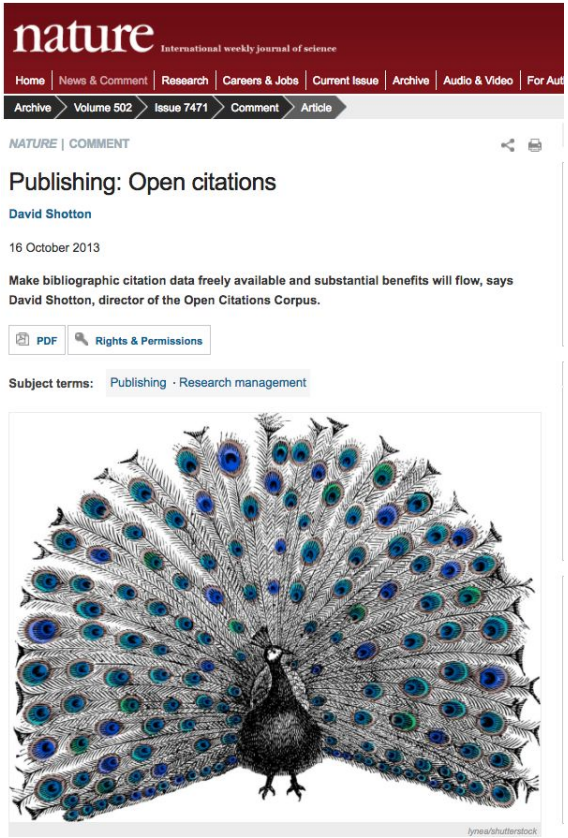
STAKEHOLDERS OF THE **INITIATIVE FOR OPEN CITATIONS** • https://i4oc.org/#stakeholders

Data reuse

The Open Citations Corpus

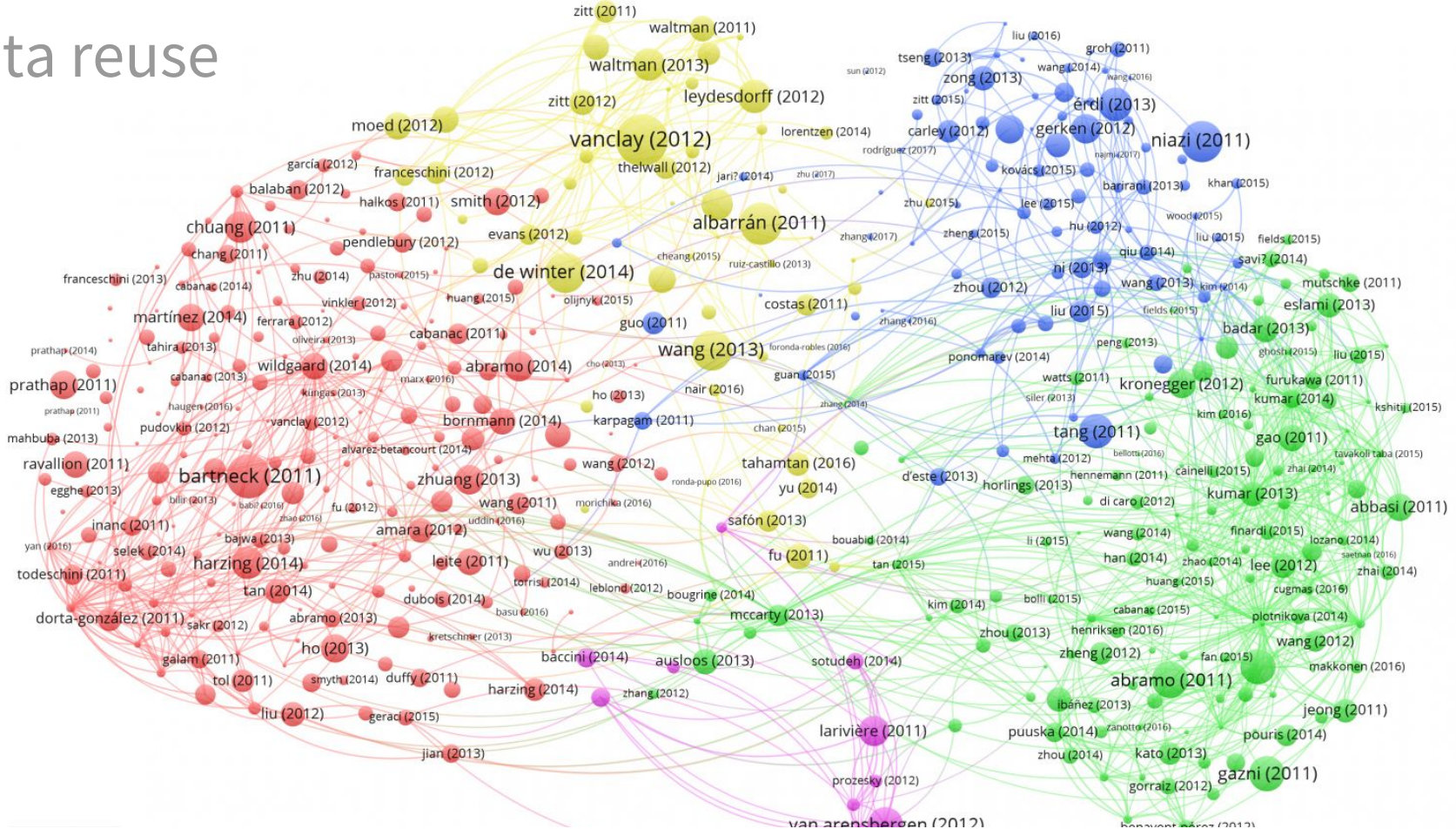
A broad and open collection of citation information from many sources

David Shotton and Silvio Peroni



THE **OPEN CITATIONS CORPUS** · http://opencitations.net/corpus

Data reuse

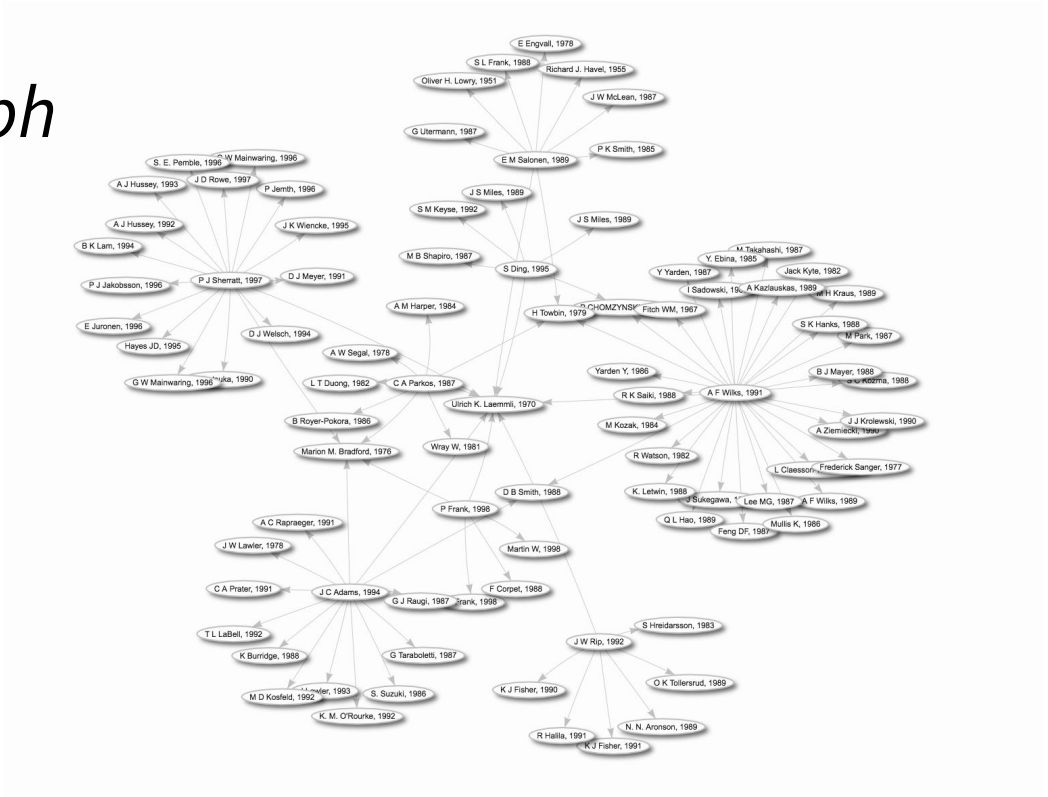


VISUALIZING FREELY AVAILABLE CITATION DATA USING **VOSVIEWER** • https://www.cwts.nl/blog?article=n-r2r294

Data reuse

The Wikidata Citation Graph

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

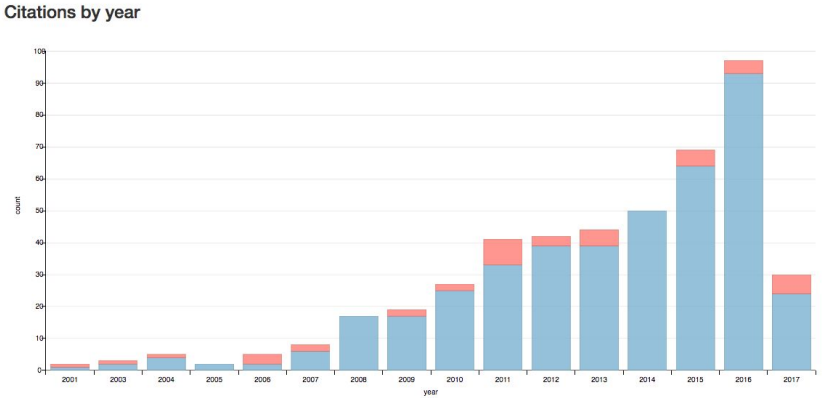


PARTIAL CITATION GRAPH FOR **ULRICH K. LAEMMLI (1970)** • http://tinyurl.com/y7acpqzd

Data reuse

Tools to create profiles

Scholia uses data from Wikidata



Citing authors

Authors that cite the author (excluding self citations).

Show 10 entries Search:

Count	Citing author
25	Christoph Steinbeck
24	Antony John Williams
21	Sean Ekins
17	Alex M Clark
13	Peter Murray-Rust
12	Ola Spjuth
11	John P. Overington
11	Nina Jeliazkova
10	Oliver Fiehn
10	Tobias Kind

Edit on query.Wikidata.org

PROFILE INFORMATION FOR **EGON WILLIGHAGEN** • https://tools.wmflabs.org/scholia/author/Q20895241

The road ahead

Lessons learned

A single, measurable goal

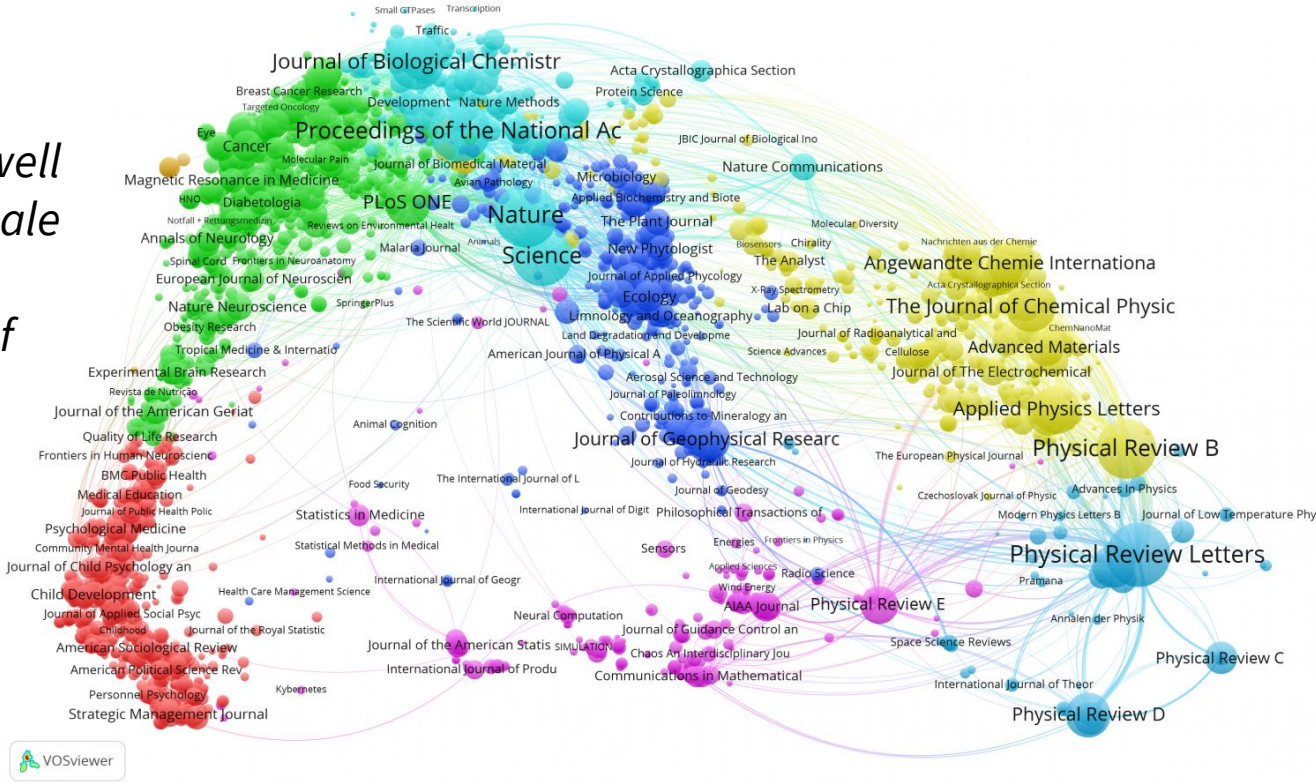
Low cost

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.

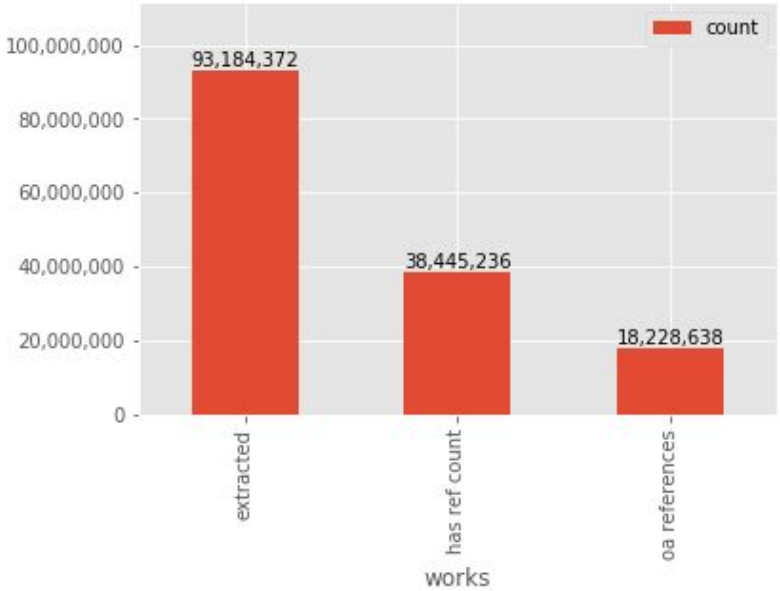
- Authors will have consistent, machine-readable access to references for all their publications;
- Researchers will be able to use this resource to study the dissemination of methods and scientific ideas, the genesis and provenance of scholarly knowledge;
- Funders will be able to rely on a public resource 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.

OPENING UP RESEARCH CITATIONS: A Q&A WITH DARIO TARABORELLI • http://bit.ly/2hfnC3b

Challenges: coverage

41% Crossref records have reference data

47% of those have open reference data



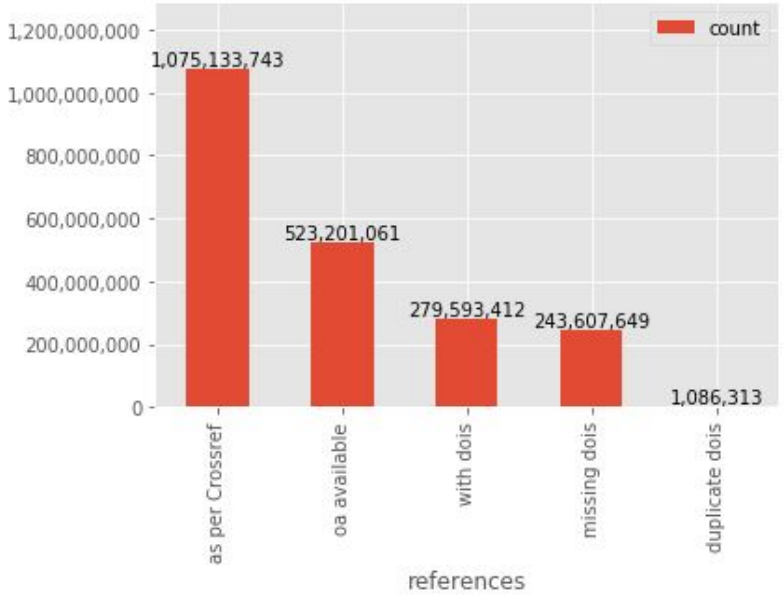
Acknowledgment: Daniel Ecer, Data Scientist, eLife. See https://elifesci.org/crossref-data-notebook

Challenges: data quality

Over 1 billion references

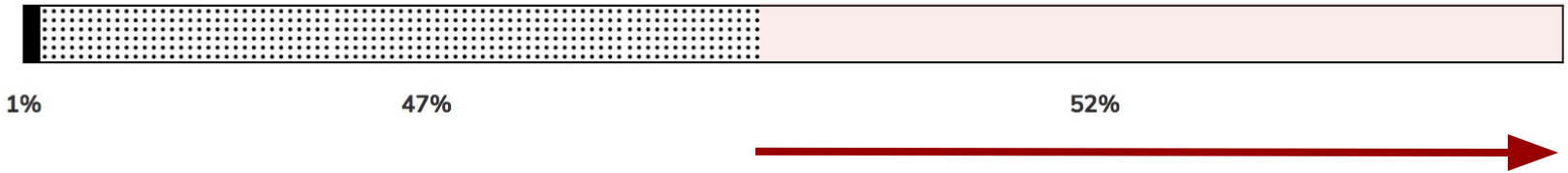
49% are open

53% have DOIs (and can be linked to another record)



Acknowledgment: Daniel Ecer, Data Scientist, eLife. See https://elifesci.org/crossref-data-notebook

The road to 100%




WILEY


Taylor & Francis Group
an informa business


ROYAL SOCIETY
OF CHEMISTRY

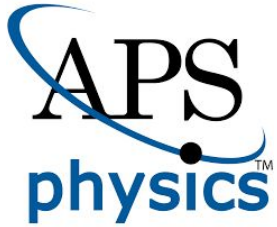

CAMBRIDGE
UNIVERSITY PRESS


SPRINGER
NATURE


SAGE
Publishing


PLOS


SciELO


APS
physics


AIP
American Institute
of Physics


DE
G


DE GRUYTER


AIAA
American Institute of
Aeronautics and Astronautics


emerald
PUBLISHING

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

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

CROSSREF MEMBERS WITH OPEN REFERENCES • https://www.crossref.org/reports/members-with-open-references/

The road to 100%

"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.

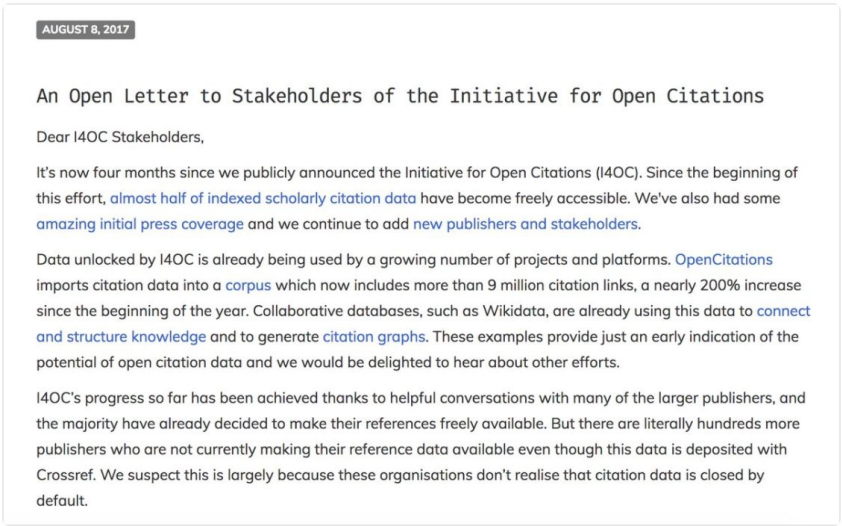
OPEN CITATIONS: A LETTER FROM THE SCIENTOMETRIC COMMUNITY TO SCHOLARLY PUBLISHERS
http://issi-society.org/open-citations-letter

Getting involved

I4OC **I4OC**
@i4oc_org

Following

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



AUGUST 8, 2017

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

123 Retweets **99** Likes

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

Thank you

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

SWIB 2017 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) *and participating publishers* (i4oc.org/#publishers)