



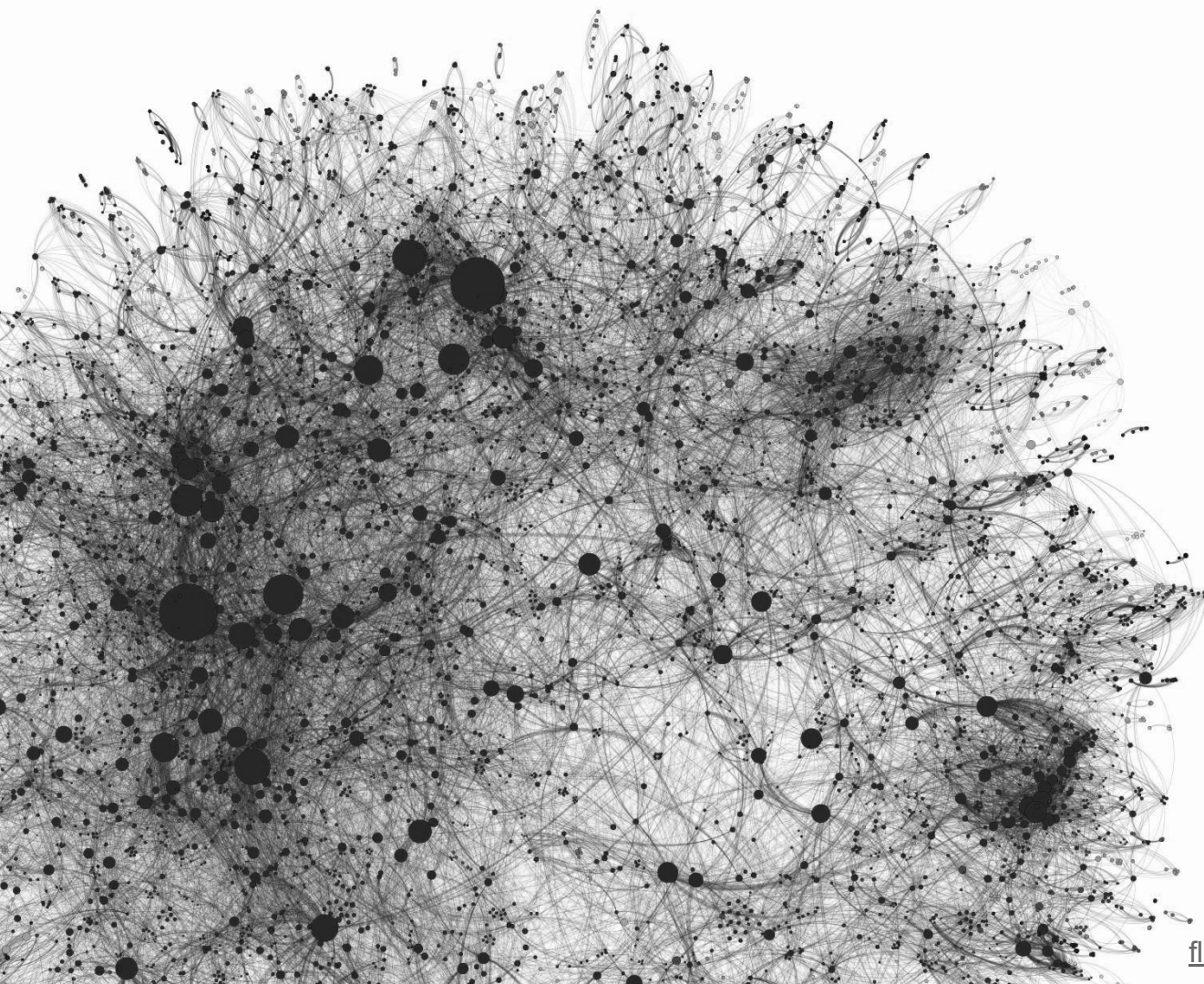
# Remixing the graph

Dario Taraborelli

Workshop on Open Citations • Bologna, 3 September 2018



why the citation graph  
should belong to the public



provenance

ANDY LAMB [CC BY]

[flickr.com/photos/speedoflife/8273922515](https://www.flickr.com/photos/speedoflife/8273922515)



impact



funding







# proprietary citation indexes

Scopus<sup>®</sup>

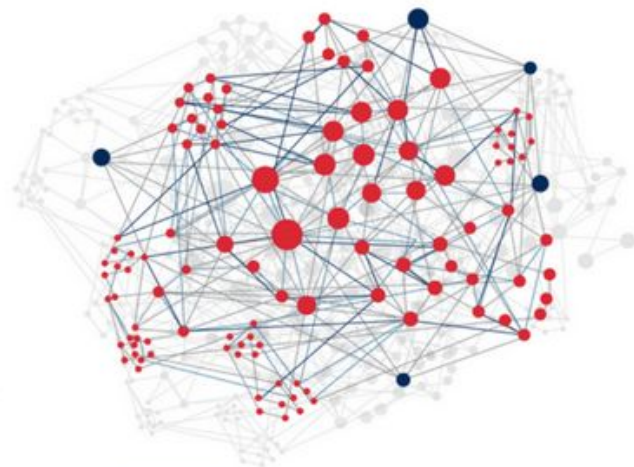
Scopus is the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings.

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

*We are **independent and unbiased**.*



# semi-proprietary citation indexes



Semantic **Scholar**



Microsoft Academic Graph





**WikiCite**

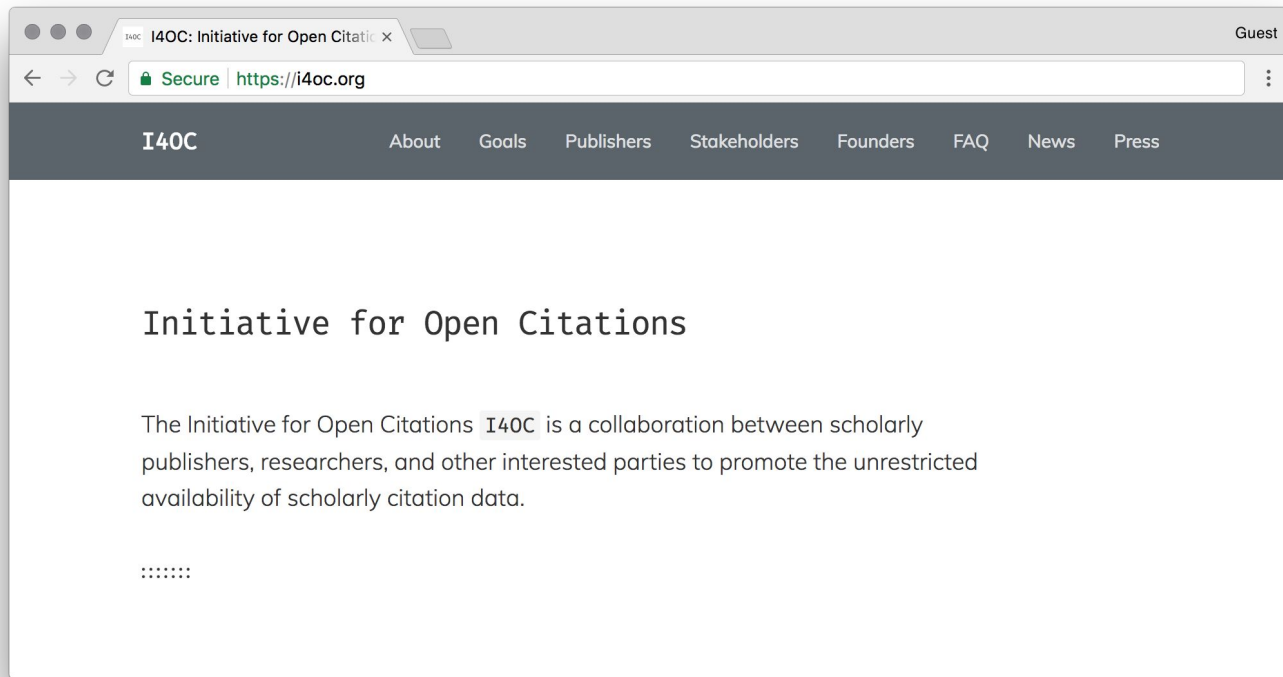
@Wikicite

"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



# 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

## The starting point

Most publishers already deposit their reference data with Crossref

The default state for the data was *closed*

## The challenge

Could we persuade a group of influential publishers to release their data all at once?



Following



Out of 999 scholarly publishers depositing reference data to **@CrossrefOrg**, only 28 (3%) are making them open

[docs.google.com/spreadsheets/d ...](https://docs.google.com/spreadsheets/d/...)  
**#COASP8**



### (ARCHIVED) Publishers depositing citation data to Crossref...

Publisher list depositing open references PUBLISHERS\*,\* publishers listed are depositing references for at least one journal. they may not be doing so for all titles. American Geophysical...

[docs.google.com](https://docs.google.com)

8:37 AM - 22 Sep 2016



# 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](mailto: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

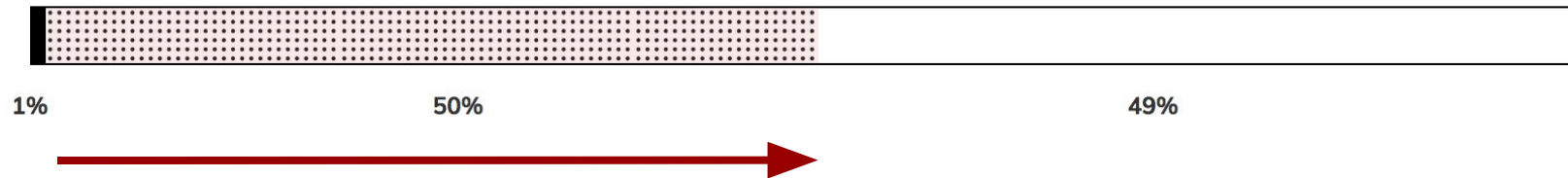
How many citations are open today?



1%

# Progress

How many citations are open today?



Progress

**20 million**

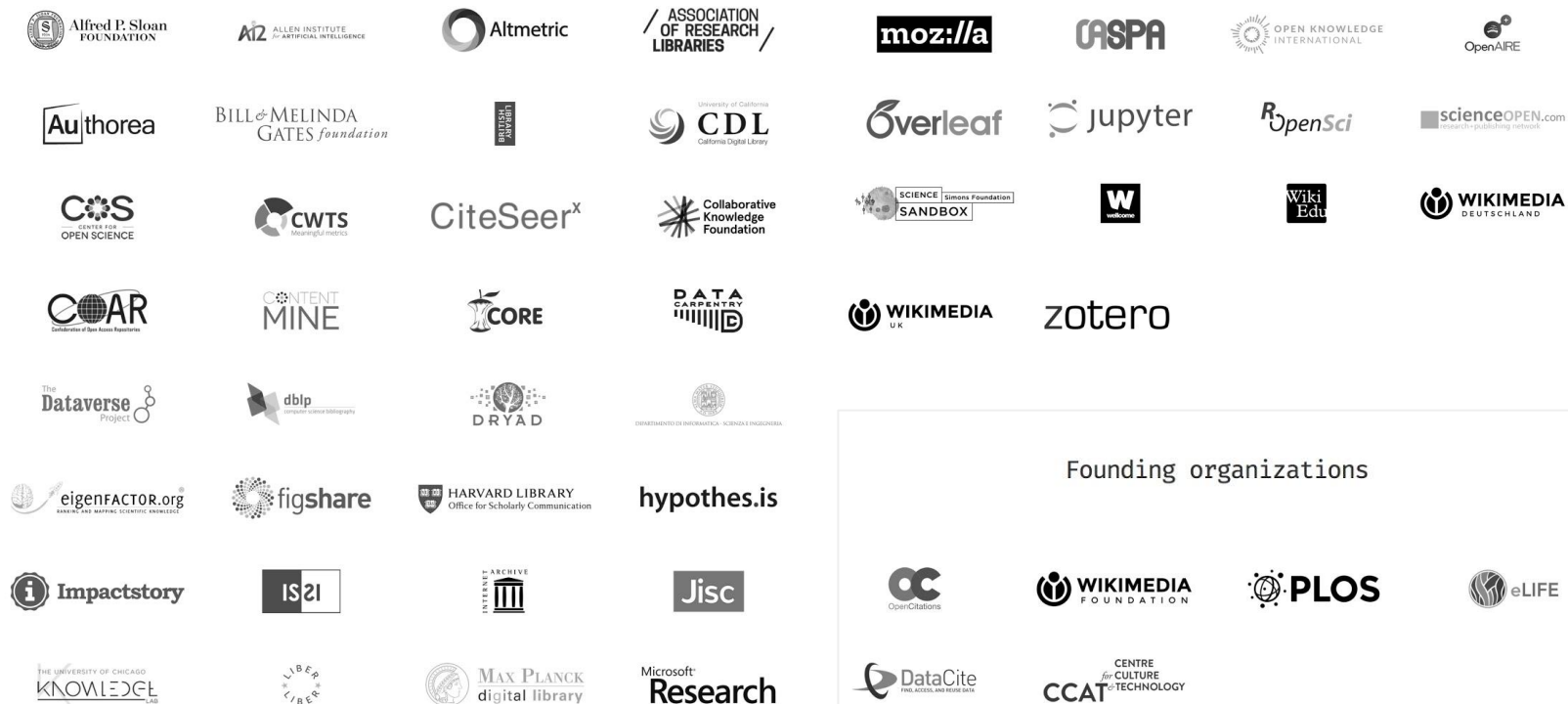
DOI records with open references

Progress

**584 million**

open reference data points

# Stakeholders



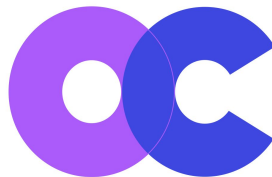


# Stakeholders



Data reuse

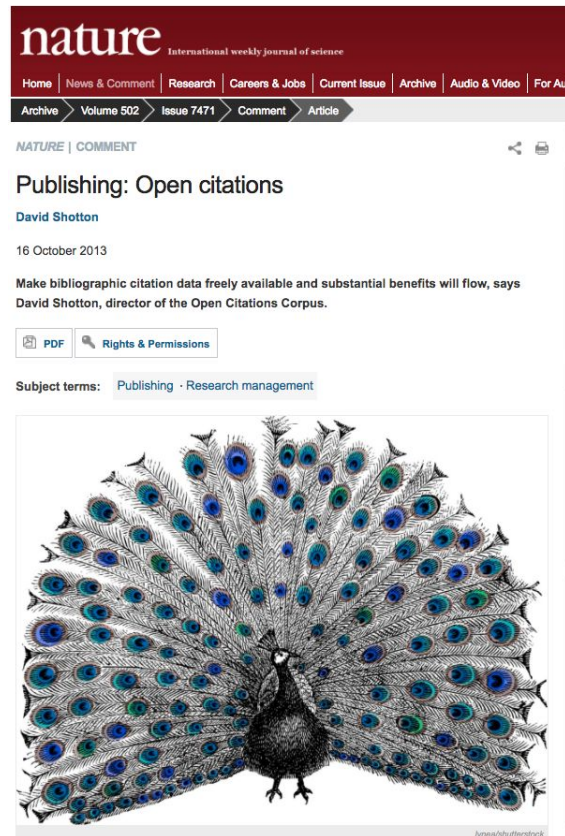
## *The Open Citations Corpus*



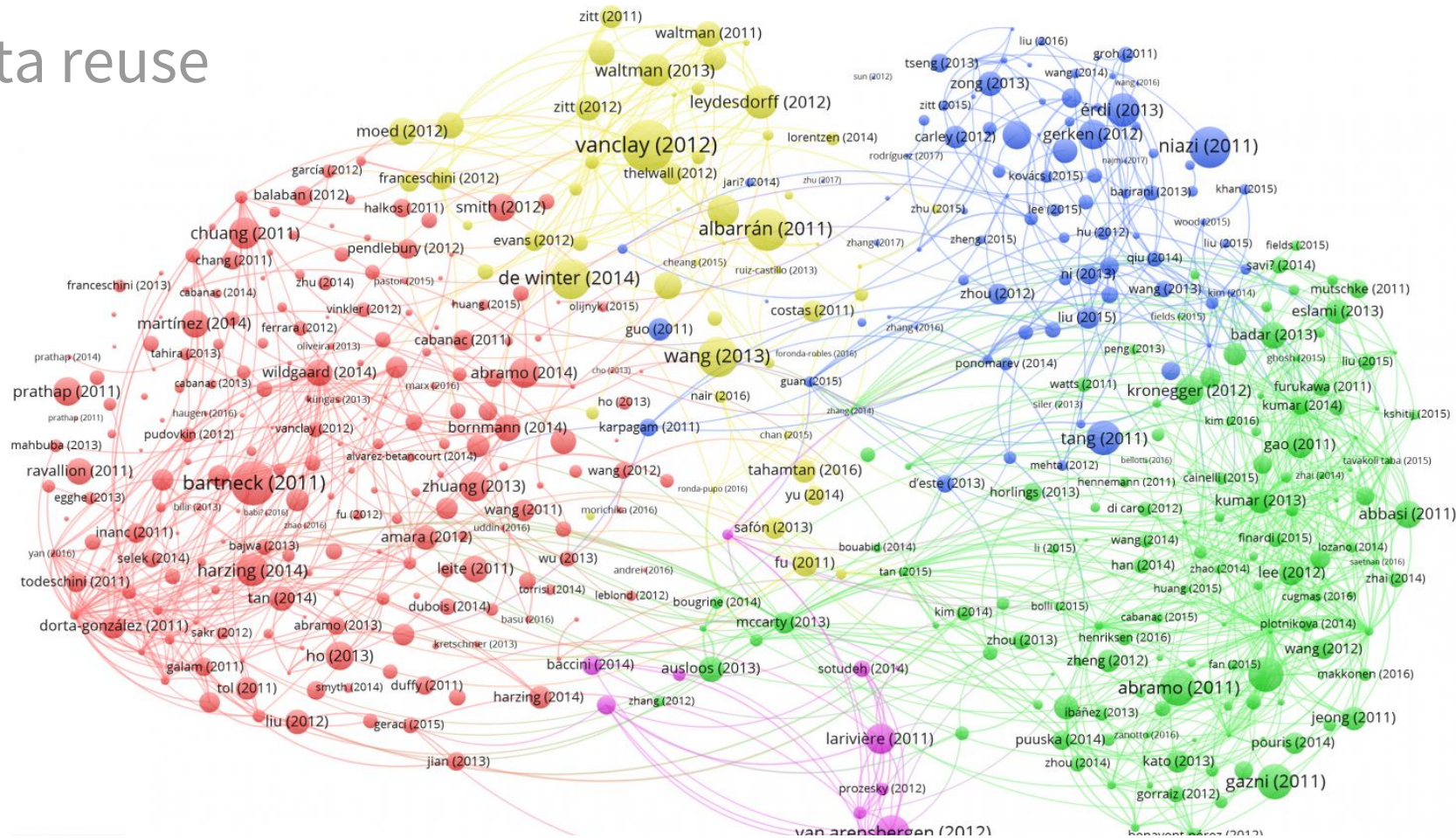
A broad and open collection of citation information from many sources

David Shotton and Silvio Peroni

THE OPEN CITATIONS CORPUS • <http://opencitations.net/corpus>



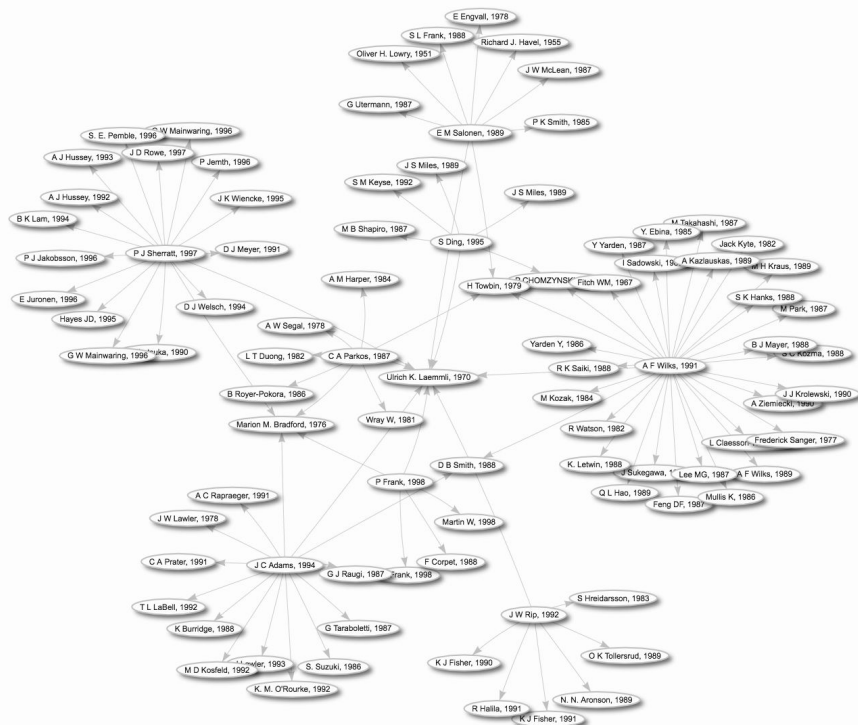
# Data reuse



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

# The Wikidata Citation Graph

using the *cites* (P2860)  
property in Wikidata

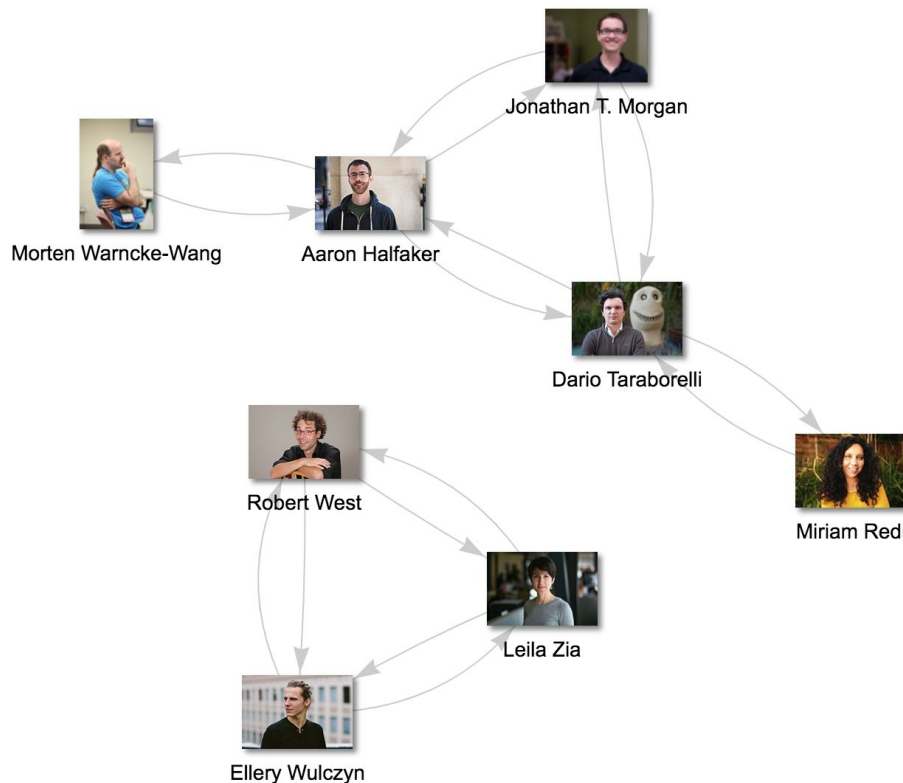




# Data reuse

## *The Wikidata Citation Graph*

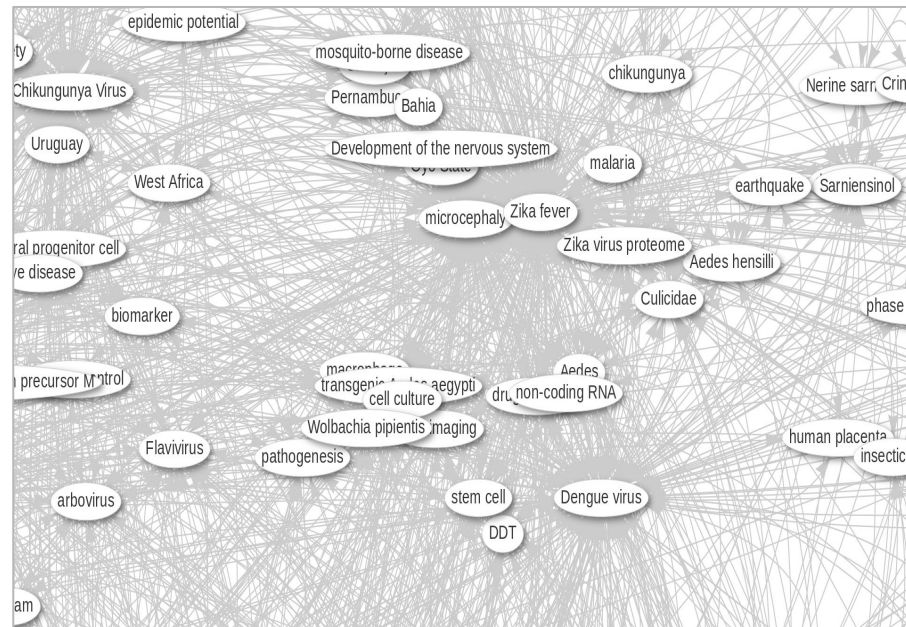
**co-citation / co-authorship**  
relations in Wikidata



Data reuse

## *The Wikidata Citation Graph*

the complete, **annotated**  
**scholarly publication** corpus  
on Zika virus (Q202864)



# Data reuse

Identify Wikidata statements or Wikipedia sentences:

citing journal articles by *physicists at Oxford University in the 1970s*

citing *a journal article that was retracted*

*lacking citations* from sources in a local language

Identify citations to a scholarly work:

by *female biologists*

by *scientists whose PhD is in a different field than the field of the cited work*

by *authors whose work was funded by a private funder*

by *authors born in the same country as the main topic of the cited work*

SOURCE-RELATED SPARQL QUERIES SUPPORTED IN WIKIDATA

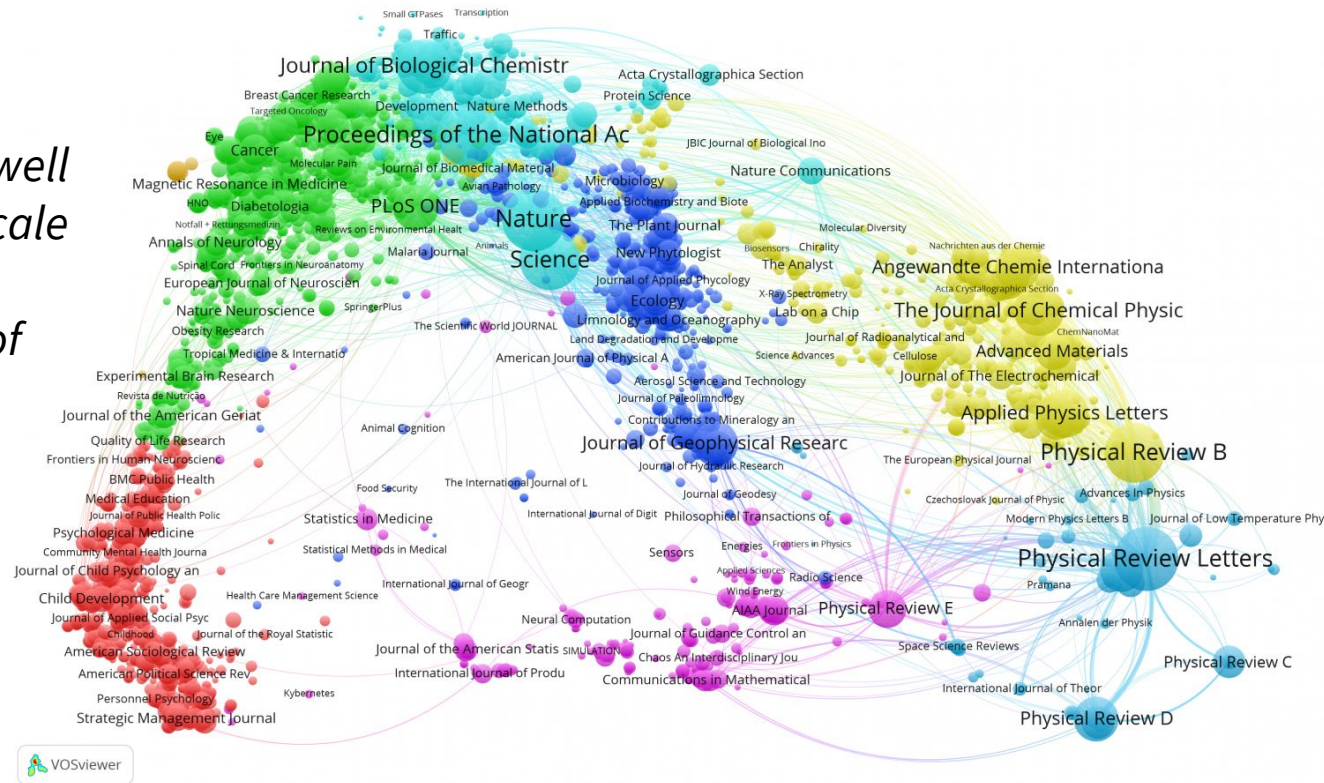
[meta.wikimedia.org/wiki/WikiCite\\_2016/Report/Group\\_5](https://meta.wikimedia.org/wiki/WikiCite_2016/Report/Group_5)



The road ahead

# 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.”*



# Who benefits from this

- **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.



advocating for open citations

# Data quality and coverage

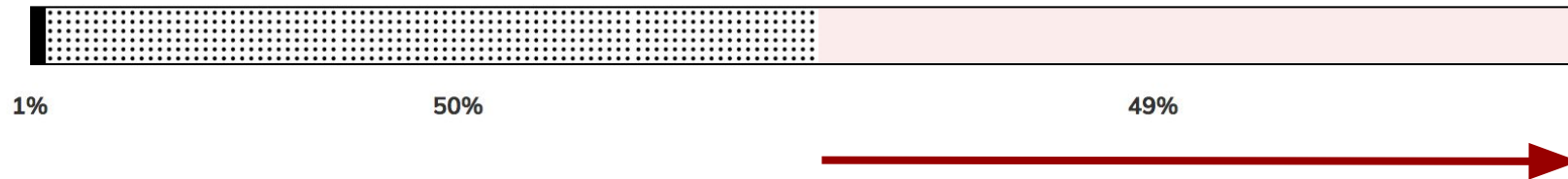
Over 1.1 billion references in Crossref

Only 56% of open references have DOIs  
(and can be linked to another record)

*what about other citations?*

# Data quality and coverage

How many citations are open today?



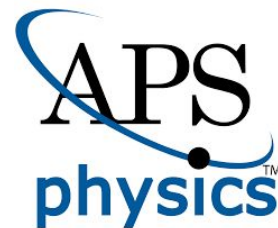


Taylor & Francis Group  
an **informa** business



CAMBRIDGE  
UNIVERSITY PRESS

**SPRINGER**  
**NATURE**



DE GRUYTER



# The road to 100%

Largest publishers among  
the top 20 DOI depositors *not*  
*distributing open references*  
(as of September 2018)

- ⊖ Elsevier
- ⊖ IEEE
- ⊖ Wolters Kluwer Health
- ⊖ IOP Publishing
- ✓ ~~Oxford University Press~~
- ⊖ American Chemical Society





building on open citations

# { } wikicite



Berkeley, 27-29 November 2018

Bay Bridge by Tehani Schroeder • [flic.kr/p/oFz47p](https://flic.kr/p/oFz47p) • CC BY

# Thank you

D. Taraborelli (2018) *Remixing the graph*.

Workshop on Open Citations, Bologna, 3 September 2018 [CC BY 4.0]

[doi.org/10.6084/m9.figshare.7038695](https://doi.org/10.6084/m9.figshare.7038695)

## Acknowledgments

The I4OC founders ([i4oc.org/#founders](https://i4oc.org/#founders)) • the I4OC stakeholders ([i4oc.org/#stakeholders](https://i4oc.org/#stakeholders)) and participating publishers ([i4oc.org/#publishers](https://i4oc.org/#publishers)) • Daniel Ecer for data analysis of the Crossref corpus • the WikiCite and Wikidata community ([meta.wikimedia.org/wiki/WikiCite](https://meta.wikimedia.org/wiki/WikiCite)).

## Additional image credits

*Light* by Numero Uno • *Fist* by Bouwe van der Molen • *Tools* by Alex Sauda Samora

**CC BY 4.0** images from the Noun Project