Open Metrics

Silvio Peroni
silvio.peroni@unibo.it – https://orcid.org/0000-0003-0530-4305 – @essepuntato

Open Science (A.Y. 2023/2024)
Second Cycle Degree in Digital Humanities and Digital Knowledge
Alma Mater Studiorum - Università di Bologna

CC BY

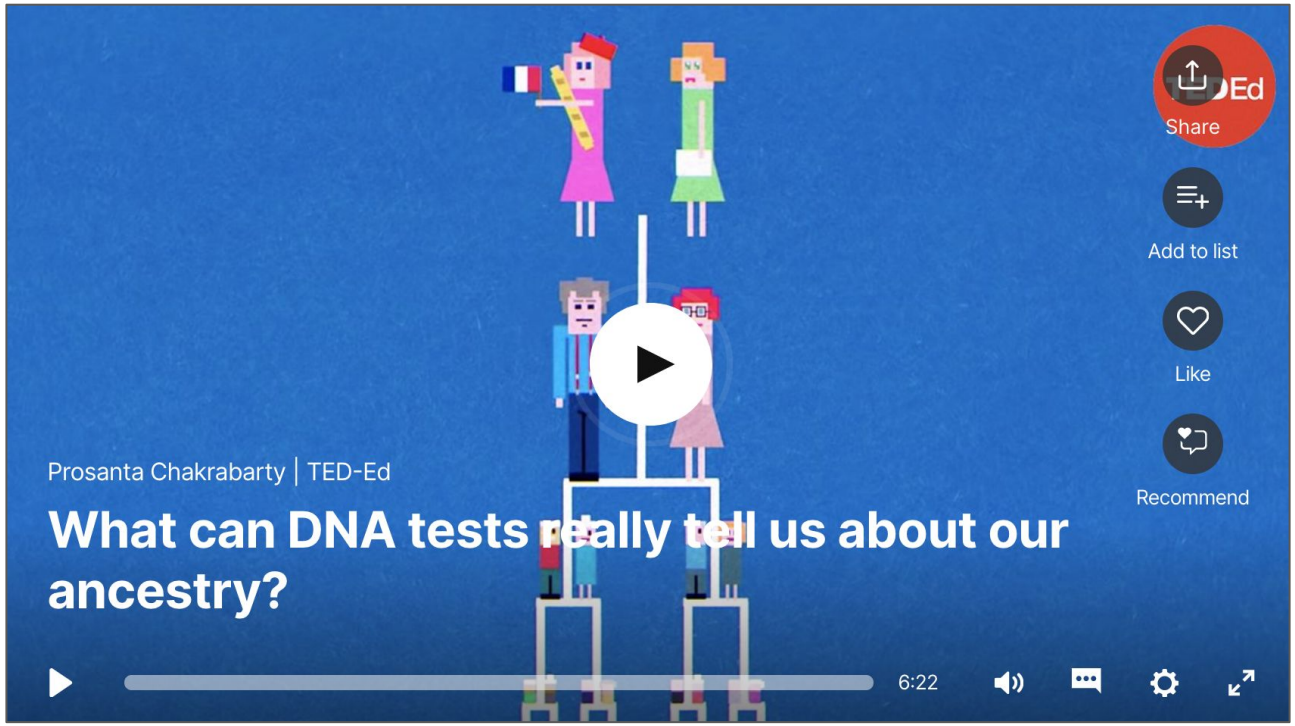
DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA

Research workflow

We discuss about the methods that are used to assess research, with a special focus on well-know bibliometrics and the importance of enabling transparency in research assessment exercises

Kramer, B., & Bosman, J. (2015, June 18). The good, the efficient and the open—Changing research workflows and the need to move from Open Access to Open Science. CERN Workshop on Innovations in Scholarly Communication (OAI9), University of Geneva, Geneva, Switzerland. https://www.slideshare.net/bmkramer/the-good-the-efficient-and-the-open-oai9

An introductory video about DNA and ancestry



Artrake Studio. (2020, June). What can DNA test really tell us about our ancestry? https://www.ted.com/talks/prosanta_chakrabarty_what_can_dna_tests_really_tell_us_about_our_ancestry

Origin of all sins

“A citation index to science would have [...] articles that had referred to the article in question, together with an indication of whether the citing source was an original article, review, abstract, review article, patent, or translation, and so forth.

[...]

Thus, in the case of a highly significant article, the citation index has a quantitative value, for it may help the historian to measure the influence of the article – that is, its ‘impact factor’”

$$\text{Impact Factor of journal J in year Y} = \frac{\text{Citations in year Y to publications published in J between Y-1 and Y-2}}{\text{Number of publications published in J between Y-1 and Y-2}}$$

Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. Science, 122(3159), 108–111. https://doi.org/10.1126/science.122.3159.108, available in OA at http://garfield.library.upenn.edu/essays/v6p468y1983.pdf

Impact Factor (IF)

The Impact Factor of a journal measures the yearly average number of citations to recent articles (i.e. published in the past two years) in that journal

The IF was developed for helping librarians to select additional source journals to be included in a library catalogue – i.e. it was a tool to help librarians identify journals to purchase

The IF for journal is a calculated by Clarivate Analytics out of the citation data included in Web of Science (WoS), one of the most famous and proprietary citation indexes

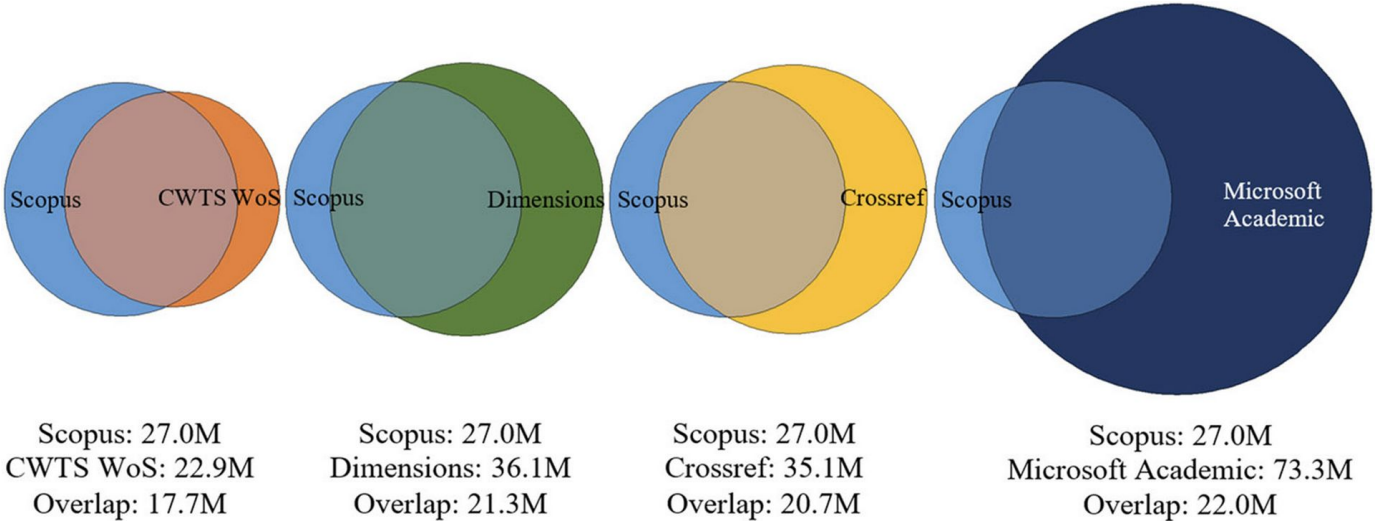
Even if the IF is a measure associated to journals, it is and has been used to measure the impact of individuals (i.e. authors) and institutions

According to Eugene Garfield, its creator, the “use of journal impacts in evaluating individuals has its inherent dangers [since] in an ideal world, evaluators would read each article and make personal judgments”

Garfield, E. (2006). The History and Meaning of the Journal Impact Factor. JAMA, 295(1), 90. https://doi.org/10.1001/jama.295.1.90, available in OA at https://garfield.library.upenn.edu/papers/jamajif2006.pdf

Is WoS fair concerning research assessment?

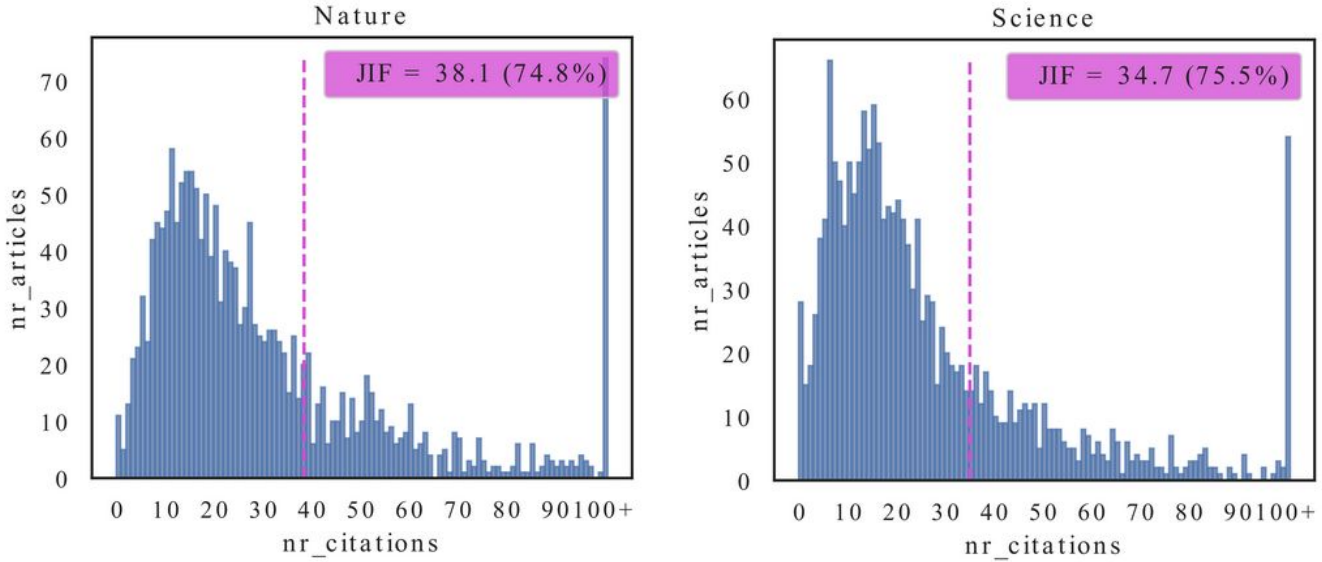
The figure shows a comparison between four distinct citation indexes, Scopus (commercial, Elsevier), WoS (commercial, Clarivate Analytics), Crossref (public, Crossref), Dimensions (private, Digital Science), Microsoft Academic (public, Microsoft)



Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quantitative Science Studies, 2(1), 20–41. https://doi.org/10.1162/qss_a_00112

Intrinsic issues of IF

“The distributions are clearly skewed, making the arithmetic mean an inappropriate statistic to use to say anything about individual papers”



Citation distributions follows the Pareto principle: **75% of citations** comes from **25% of articles**

Tennant, J. P., Crane, H., Crick, T., Davila, J., Enkhbayar, A., Havemann, J., Kramer, B., Martin, R., Masuzzo, P., Nobes, A., Rice, C., Rivera-López, B., Ross-Hellauer, T., Sattler, S., Thacker, P. D., & Vanholsbeeck, M. (2019). Ten Hot Topics around Scholarly Publishing. Publications, 7(2), 34. https://doi.org/10.3390/publications7020034

DORA

In 2012, during the Annual Meeting of The American Society for Cell Biology in San Francisco (US), a group of editors and scholarly publishers discussed about the issues related to how the quality of research output is evaluated, and how the primary scientific literature is cited

In particular, they believed that impact factors for journals do not accurately reflect the value to the community of the work published in these journals, and came out with a set of recommendations referred to as the San Francisco Declaration on Research Assessment (DORA), published in May 2013

The current version of the declaration is composed by 18 guidelines composed by a general one plus others having as primary audience funding agencies, institutions, publishers, suppliers of metrics, and researchers

General recommendation: do not use journal-based metrics (e.g. IF) as a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions, or in hiring, promotion, funding decisions

17,235 individuals and 2,199 organizations in 145 countries have signed DORA as of 21 April 2021

Cagan, R. (2013). The San Francisco Declaration on Research Assessment. Disease Models & Mechanisms, 6(4), 869–870. https://doi.org/10.1242/dmm.012955

What to use as proxy for quality of articles

Data are increasingly used to govern science, and research evaluations rely on metrics: evaluation is led by the data rather than by judgement, and metrics have proliferated in the past years – usually well intentioned, not always well informed, often ill applied, thus risking to damage the system with the very tools designed to improve it

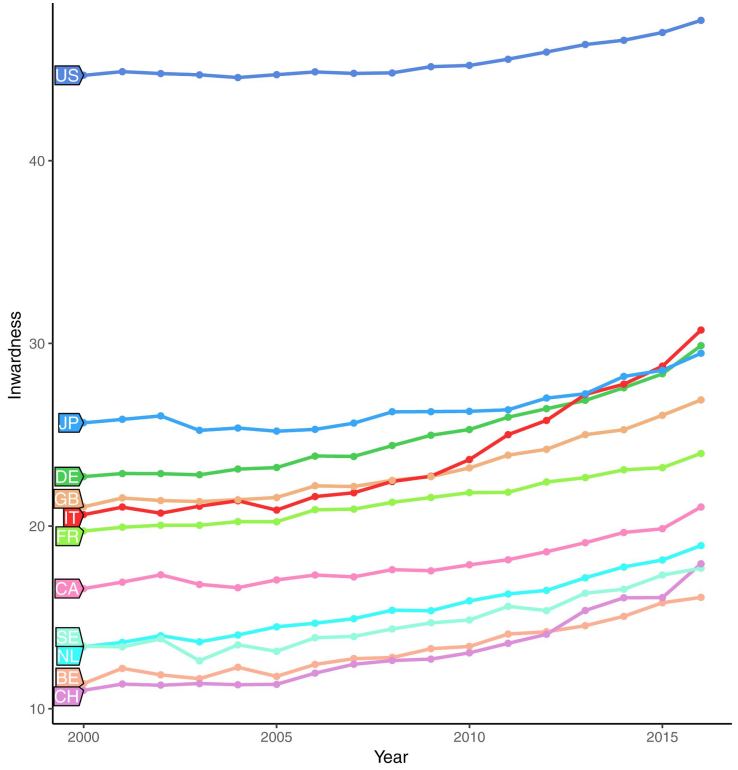
In 2014, a group of scientometricians launched the Leiden Manifesto for Research Metrics, which comprises ten principles to guide research evaluation, which has been also summarised in a video summary

Among these guidelines, the suggestion is to use citation count as a proxy for the quality of an article, since it explicitly refers to it, keeping into consideration, though, possible issue that may arise even in this context:

- The time in which it is accumulated vary depending on several factors, such as its research discipline and when and how (OA vs. non-OA) the article was available for the first time on the Web (preprint, postprint, etc.)
- Its use as a single indicator for evaluating the quality of a research will invite gaming (Goodhart's law: when a measure becomes a target, it ceases to be a good measure) – thus, it is suggested to use a suite of indicators to provide a more robust and pluralistic picture

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. Nature, 520(7548), 429–431. https://doi.org/10.1038/520429a

A example of gaming based on citation counts



Inwardness: the proportion of citations coming from a country over the total number of citations gathered by that country

The change in the Italian trend occurred in the years following the introduction in 2011 of national regulations in which key passages of professional careers are governed by bibliometric indicators – and a generalized strategic use of citations in the Italian scientific community was observed, both in the form of author self-citations and of citation clubs

Baccini, A., De Nicolao, G., & Petrovich, E. (2019). Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. PLOS ONE, 14(9), e0221212. https://doi.org/10.1371/journal.pone.0221212, available in OA at https://arxiv.org/abs/1801.08992

Evaluating new journals and new articles

Even if one institution continues to pursue the adoption of the IF as a proxy of the quality of a journal, which strategy one can use to assess the quality of new journals if no IF is available, either when the journal is not listed in WoS or when it is a new journal?

In some cases, there are other metrics available that have been released by different companies and institutions, such as the Eigenfactor Metrics, the Source Normalized Impact per Paper (SNIP), the CiteScore, and the SCImago Journal Rank (SJR)

About articles, which strategy can be adopted to assess the quality of new articles if no citation count is available – since it is reasonable to see the first citations only after a few months from the official publication date?

Larivière, V., & Sugimoto, C. R. (2019). The Journal Impact Factor: A Brief History, Critique, and Discussion of Adverse Effects. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), Springer Handbook of Science and Technology Indicators (pp. 3–24). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_1

Altmetrics

Altmetrics is the study and use of scholarly impact measures based on activity in online tools and environments

Usually, they are are fast to accumulate, and they can be gathered via public APIs

They comprise at least four distinct dimensions:

1. Views: HTML views and PDF downloads
2. Discussion: journal comments, science blogs, Twitter, Facebook and other social media such as Scholarly Social Networks (Research Gate, Academia.edu)
3. Saving: Mendeley, Zotero and other social reference managers
4. Non-academic citations: references in Wikipedia

Unlike citation metrics, altmetrics track impact outside the academia

Priem, J., Groth, P., & Taraborelli, D. (2012). The Altmetrics Collection. PLoS ONE, 7(11), e48753. https://doi.org/10.1371/journal.pone.0048753

Reproducibility crisis in research assessment

The reproducibility crisis does not affect only science, but also the methods we currently use to assess it

Not only a critique to the ‘recipes’ (i.e. the metrics) used in research assessment, but also to the lack of transparency of the ‘ingredients’ (i.e. the data) used to compute recipes outcomes

It should always be possible to dig deeper into the data and to see what is behind a certain number; this requires scholarly metadata and citation data to be open rather than paywalled and accessible only by paying a fee – as in the case of commercial citations indexes, such as WoS and Scopus, which are the two most adopted databases in research assessment exercises

There is an urgent need of a global community effort in the scholarly domain to put such ingredients to the commons – often they are pure facts, i.e. data that cannot be copyrighted such as citations, and often they are used in metrics-based research assessment

Waltman, L. (2020, September 9). Responsible research assessment requires open scholarly metadata. Workshop on Open Citations and Open Scholarly Metadata 2020, Bologna, Italy. Zenodo. https://doi.org/10.5281/zenodo.4021492

Pushing for open data in research metrics

National Plan For Open Science, https://www.ouvrirlascience.fr/national-plan-for-open-science-4th-july-2018/

"Assessment system for researchers and research institutions must be updated to reflect the principles and practices of open science"

San Francisco Declaration on Research Assessment, https://sfdora.org

"Be open and transparent by providing data and methods used to calculate all metrics"

"Provide the data under a license that allows unrestricted reuse, and provide computational access to data"

Leiden Manifesto for Research Metrics, http://www.leidenmanifesto.org/

"Keep data collection and analytical processes open, transparent and simple"

"Allow those evaluated to verify data and analysis"

Paywalled and close citation data do not comply with the principles above – indeed such data are a threat to transparency, replicability and verifiability of research assessment exercises

Citations as common good

The concept of open citations is strongly tied with that of the Web

Since 1989, the Web has drastically change the way we think academic publishing and science in general – they started to adopt Web standards to create and deliver their products quickly and to a broader audience

Standards (e.g. JATS), guidelines (e.g. FAIR), services (e.g. REST APIs) based on Web technologies have been proposed in the past 30 years to improve the discoverability of academic products and publications, to improve research practices and to allow reusability of scholarly data in different applicative contexts

Open citations are no exception

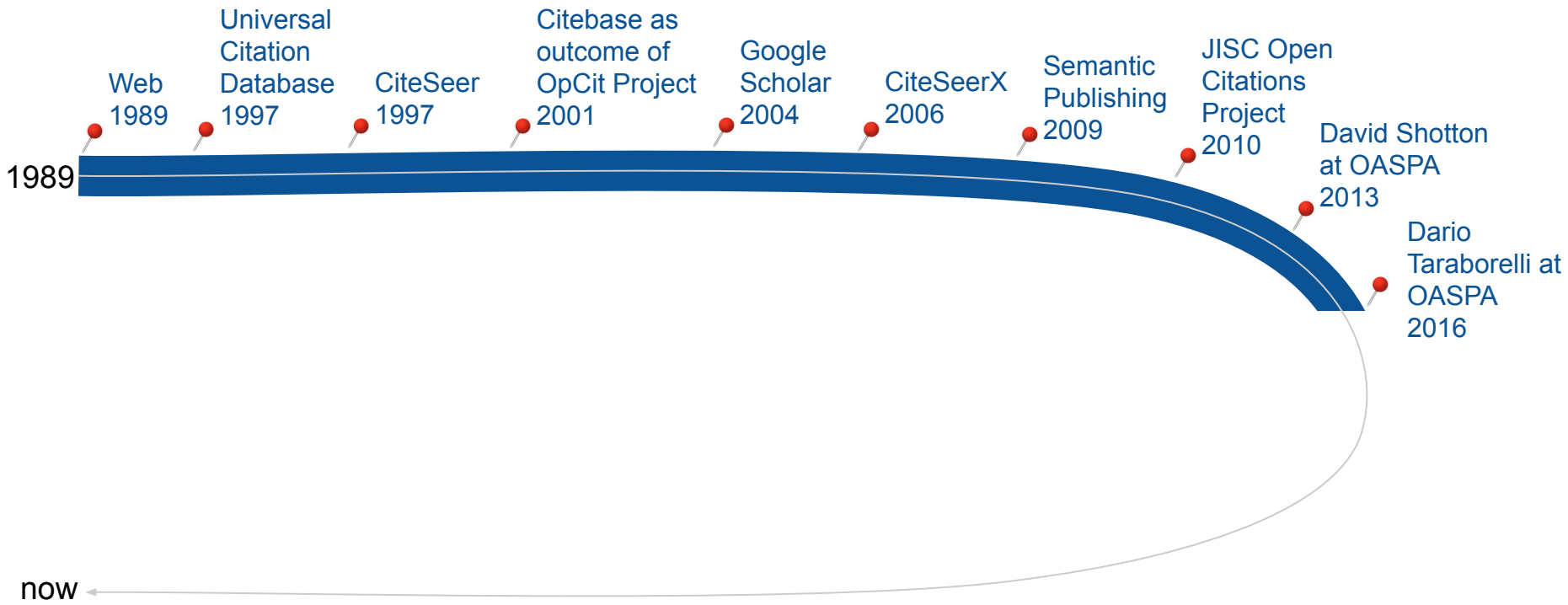
The origins

The first embryonal introduction of open citations is in Robert Cameron's visionary article published in 1997, in which he speculates about the existence of a Universal Citation Database

- Linking every scholarly work ever written
- Freely available over the Internet
- Updated every day
- Comprehensive (no selection of particular venues, all are included)
- All types of publications (from articles to working papers and preprints)
- All publications are equally visible (even if not equally accessible)
- Decentralised

Cameron, R. D. (1997). A Universal Citation Database As a Catalyst For Reform In Scholarly Communication. First Monday, 2(4). https://doi.org/10.5210/fm.v2i4.522

A timeline: the origins



Scandalous references

"In this open-access age, it is a scandal that reference lists from journal articles — core elements of scholarly communication that permit the attribution of credit and integrate our independent research endeavours — are not readily and freely available for use by all scholars"

Shotton, D. (2013). Open citations. Nature, 502(7471), 295–297. https://doi.org/10.1038/502295a

A timeline: open citations

1989

Web 1989

Universal Citation Database 1997

CiteSeer 1997

Citebase as outcome of OpCit Project 2001

Google Scholar 2004

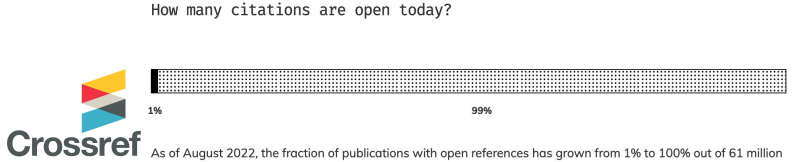
CiteSeerX 2006

Semantic Publishing 2009

JISC Open Citations Project 2010

David Shotton at OASPA 2013

Dario Taraborelli at OASPA 2016



From https://i4oc.org

Wor
Open

now

A timeline: so far

1989

Web 1989

Universal Citation Database 1997

CiteSeer 1997

Citebase as outcome of OpCit Project 2001

Google Scholar 2004

CiteSeerX 2006

Semantic Publishing 2009

JISC Open Citations Project 2010

David Shotton at OASPA 2013

Dario Taraborelli at OASPA 2016

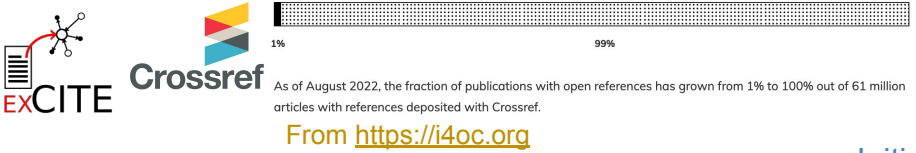
OpenCitations WikiCite 2016

Initiative for Open Citations 2017

WikiCite Workshop 2017

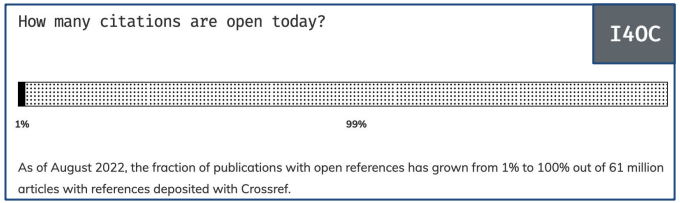
Workshop on Open Citations 2018

now



From https://i4oc.org

Everything good? Well, it is a start, but…



There are more than 13,000 additional publishers who are depositing publication metadata with Crossref, but who are failing to submit references with the other publication metadata. I4OC appeals to these publishers to start depositing their publication references, and other rich metadata including publication abstracts, ORCID author identifiers, and funder information, along with the basic metadata they already submit to Crossref.

Publishers should commit to making complete metadata for all their works, including references and abstracts, available in a suitable open infrastructures such as Crossref or DataCite

Publishers that do not yet support the Initiative for Open Citations (I4OC, https://i4oc.org) and the Initiative for Open Abstracts (I4OA, https://i4oa.org) should join these initiatives

Likewise, preprint servers and institutional repositories should also deposit their metadata to open infrastructures

Hendricks, G., Kramer, B., Maccallum, C., Manghi, P., Neylon, C., Peroni, S., Shotton, D., Tay, A., & Waltman, L. (2021, ottobre 27). Now is the time to work together toward open infrastructures for scholarly metadata. [Blog]. LSE Impact Blog. https://blogs.lse.ac.uk/impactofsocialsciences/2021/10/27/now-is-the-time-to-work-together-toward-open-infrastructures-for-scholarly-metadata/

End

Open Metrics

Silvio Peroni
silvio.peroni@unibo.it – https://orcid.org/0000-0003-0530-4305 – @essepuntato

Open Science (A.Y. 2023/2024)
Second Cycle Degree in Digital Humanities and Digital Knowledge
Alma Mater Studiorum - Università di Bologna

CC BY

DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA