null • Business of Knowing

The Invisible Citation Commons

Phoebe Ayers, Samuel J. Klein

Published on: Aug 17, 2021

DOI: 10.21428/6ffd8432.5af8c64c

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Introduction

Scholarly knowledge relies on citations. Discovering and acknowledging prior work is fundamental to knowing what has been done before, synthesizing the state of the field, and identifying spaces for new research.

Despite being so crucial, citations — the pieces of metadata that serve as references to works — are often ignored in discussions of types of open knowledge. Historically, citations and their cross-references have been laboriously collected (by hand, then by computational techniques) in bibliographic indexes. These indexes, once published in serial print volumes, are today generally offered as web-based, paywalled subscription products by scholarly entities or for-profit companies. Indexing and abstracting services are big business: Web of Science, an academic journal and proceedings index that has existed since the 1960s, can cost subscribing libraries hundreds of thousands of dollars a year. Web of Science is currently owned by Clarivate, which in 2020 had revenues of 1.2 billion USD¹ from its portfolio of analytics and intellectual property management tools that monetize the research process.

Indexes like Web of Science collect and annotate references with subject information, mine published works for their citations, and provide tools to help researchers discover and analyze those citations. Because of their subscription status, access to these indexes, much like subscription-based journals, is typically limited to affiliates of subscribing libraries. The introduction of Google Scholar in 2004 changed how a generation of researchers work, by providing easy and unpaywalled access to an interdisciplinary database of citations derived from the web. However, Google Scholar isn't transparent about its processes, doesn't provide openly licensed or downloadable data, and includes citations that are subject to missing information and poor disambiguation.

In recent years, there has been a push to openly license citation metadata to better enable large-scale analyses and discoverability of scholarly work. The “Initiative for Open Citations” (I4OC),² launched in 2017, has led the way in helping publishers share citations to their works under a public domain CC0 license. As of early 2021, over a billion citations from one scholarly article to another are collected in public domain databases, a major shift from just a few years earlier.³ These open databases provide the backbone for new discovery tools, and are used by academics training artificial intelligence tools. Open corpora like the Microsoft Academic Graph are themselves widely cited.⁴ However, Microsoft Academic Graph will be shuttered in 2021; despite

their importance, new citation projects are reliant on continued funding and support by their host, and longevity is not always guaranteed.

Wikidata and WikiCite

Wikidata is a freely licensed and editable online database of linked data, with 94 million items as of June 2021.⁵ Like its sister project Wikipedia, it has a vibrant multilingual volunteer community that develops and maintains it, and is supported by the non-profit Wikimedia Foundation. Wikidata also includes bibliographic metadata: as of June 2021, nearly 40 million items on Wikidata represented publications, accounting for 43% of all items.⁶ These are a combination of semi-automated uploads of citations from other open databases, items about notable publications that have their own Wikipedia articles, and items added manually by editors. Wikidata is also attractive for libraries, archives, and cultural institutions that want to make their metadata more openly available and reusable, and there are several ongoing projects to incorporate Wikidata into library and archival cataloging processes and connect Wikidata to new open knowledgebases.⁷

Wikidata items can also be created about the authors, institutions, publishers, journals, and ideas related to citations, which creates a rich network of queryable information. Wikidata items about publications can include identifiers from a vast number of other catalogs and indexes, such as national library catalogs and authority files. Thus, the Wikidata item [about the Origin of Species](#) links to the Wikidata items for Darwin and the concept of “natural selection,” but also includes 22 other national and international library catalog identifiers, as well as linking to the 73 Wikipedia articles in various language editions that exist about the work (and, because the book is in the public domain, the full text on Wikisource in six languages). Wikidata serves as a hub, including identifiers for the same work from, say, Project Gutenberg, the National Library of France, and the French Wikipedia, providing a way to map connections and coverage among these diverse entities.

WikiCite is a collective name for the volunteer community and projects focused on improving the representation of open bibliographic metadata on Wikidata and the other Wikimedia projects. WikiCite provides a home for participants from a broad range of geographies and professions — librarians, developers, GLAM practitioners, data modelers, ethnographers, and Wikidatians — who are interested in improving the citation practices and infrastructure for free knowledge. From 2016-2021, the WikiCite

project was funded by several grants to the Wikimedia Foundation, most recently from the Alfred P. Sloan Foundation, which supported a series of four community conferences and funding for innovative technical and outreach projects.⁸

This focus on bibliographic metadata in Wikidata has led to a rich ecosystem of tools developed by volunteers to assist in uploading, editing and analyzing these records. One such tool is Scholia,⁹ which creates visual scholarly profiles based on Wikidata records. Viewing a heavily-cited author — such as Jennifer Doudna, 2020 Nobel Prize winner for chemistry — shows the power that can come with visualizing citations to scholarly works. The [Wikidata item for Doudna](#) gives us biographical information such as awards received. Viewing Doudna's [author record in Scholia](#), however, provides a list of associated publications by year, a map and word cloud of topics, an interactive diagram of co-authors, and a list of citing authors, all based on citations in Wikidata. Scholia and related tools provide a possible open alternative for expensive and proprietary scholarly metrics tools that are currently sold by major companies like Elsevier and Clarivate.

Unsolved challenges and future directions for open citations

Unlike other open citation databases, Wikidata, like Wikipedia, relies on a dedicated and highly skilled global group of volunteer maintainers and editors. Though the Wikimedia Foundation provides a stable platform for Wikidata, with a long-term commitment to preservation and availability, stewarding this collection of data means continuing to develop and support the editor community and making it possible for new editors and entities to contribute. The openly editable model of Wikidata differs from traditional library catalogs or indexes, where editing is restricted to a small group of staff who also ensure quality and accuracy. In Wikidata, users of the data can also contribute both small fixes and large updates, but doing so requires learning complex new workflows and navigating Wikidata's culture.

There are technical challenges as well to representing citation metadata in Wikidata. Wikidata contains only a fraction of all open citations, which is only part of all possible bibliographic metadata; tools like Scholia draw on incomplete data. However, drastically expanding the number of items about publications (such as by importing the entire open citation corpus, which would double the current size of Wikidata) raises issues of scalability, both in terms of technical infrastructure and human

curation ability. An open question in the WikiCite community is whether items about publications should remain in Wikidata, or become a separate interlinked knowledgebase that could be connected to the other Wikimedia projects. Starting a new initiative like this is a complex decision with both technical and social implications.^{[10](#)}

A related problem is how to make citations easily reusable within the Wikimedia projects and beyond. Citations form the backbone of Wikipedia articles. In an open collaborative environment where authorship is largely pseudonymous, Wikipedia articles rely on outside references for every factual claim. However, it is not yet possible to, for instance, easily add a reference to an article in Wikipedia and see how that reference is also used in other articles and language editions, trace a citation to a retracted article, or see whether the usage of a particular citation can be characterized.^{[11](#)} The infrastructure provided by Wikidata, or by a new interlinked project focused only on bibliographic metadata, could make this possible. There are at least 29 million citations in the English Wikipedia alone.^{[12](#)} Storing the citations that Wikipedia articles across 300 languages rely on as structured data would make them available for analysis and querying, which could help identify content gaps, fight misinformation, and lead to a much deeper understanding of “how we know what we know” on Wikipedia.

Another set of challenges faced by the WikiCite community is how to represent diverse types of publications, works and authors accurately across many domains. In common with other bibliographic catalogs, the WikiCite community has struggled with how to best represent works with several versions, like books with many published editions or articles with preprints. However, Wikidata’s coverage goes beyond books and articles. It is global, and aims to represent the totality of types of human knowledge. The WikiCite community has many projects that model a range of bibliographic works in Wikidata, from Brazilian laws^{[13](#)} to palm-leaf manuscripts from Indonesia.^{[14](#)} Wikidata excels at representing multilingual data, supporting labels in many languages for each item. But much remains to be done to fully represent the diversity of citable works. Individual scholars can aid in this effort by modeling citations in their own domains on Wikidata,^{[15](#)} and ensuring that works like software, datasets, models, reports, webpages and more are accurately and comprehensively cited. Publishing platforms can also support making their works more easily citable; for instance, after years of work by researchers in the software citation community, GitHub, the popular software repository, just announced built-in citation support in July 2021.^{[16](#)}

Care must also be taken to ensure accuracy, not only with citation data but also with the accuracy of linked items, including items about authors and works. In particular, there are issues with describing authors of publications, including name disambiguation, name representation, and cleaning up incomplete or duplicative ORCID data. For items describing people, there is the potential for misrepresenting identity and background in ways that cause harm, such as with statements about gender or ethnicity. New open services are needed for disambiguating names and reconciling multiple sources. Here too publishers and cataloging bodies must drive adoption of open tools like ORCID and name authority standards that don't make assumptions about personal characteristics.

Wikidata demonstrates that data curation can work across linguistic and disciplinary boundaries. Wikidata and WikiCite shows the power of a community-maintained linked open citation graph and offers, in Meg Wach's words, "the promise of sustainable infrastructure that is informed by library values and scholar need."¹⁷ The growth of open citations will call for new technical and community innovations, such as infrastructure to interlink library catalogs with Wikidata. This will involve collaboration among volunteer and professional catalogers, indexers, technologists and researchers, working together to maintain a citation commons.

Thanks to [Daniel Mietchen](#) for help outlining this article, and to the entire WikiCite community.

□ [read more essays in the "Business of knowing" summer series](#) □

Footnotes

1. Clarivate. (Feb. 25, 2021). Clarivate reports Fourth Quarter and Full Year 2020 Results. <https://clarivate.com/news/clarivate-reports-fourth-quarter-and-full-year->

[2020-results/](#) [↵](#)

2. Initiative for Open Citations: <https://i4oc.org/> [↵](#)
3. Hutchins, B. I. (2021). A tipping point for open citation data. Quantitative Science Studies . doi: https://doi.org/10.1162/qss_c_00138 [↵](#)
4. Tay, A., Martin-Martin, A. and Hug, S.E.. (May 21, 2021). Goodbye, Microsoft Academic -- Hello, Open Research Infrastructure? LSE Impact Blog.
<https://blogs.lse.ac.uk/impactofsocialsciences/2021/05/27/goodbye-microsoft-academic-hello-open-research-infrastructure/> [↵](#)
5. Wikidata: <https://wikidata.org> [↵](#)
6. WikiCite statistics. (July 2021). <http://wikicite.org/statistics.html> [↵](#)
7. See for instance IFLA discussion series of videos about libraries and Wikidata at <https://www.youtube.com/playlist?list=PLV81siTMahbsjakNDIdFwbIvdyNgwRNPo>, and the Linked Data in Libraries (LD4) Wikidata affinity group: https://www.wikidata.org/wiki/Wikidata:WikiProject_LD4_Wikidata_Affinity_Group [↵](#)
8. WikiCite. (2021, July 14). Meta, discussion about Wikimedia projects.
<https://meta.wikimedia.org/w/index.php?title=WikiCite&oldid=21733273> . [↵](#)
9. Scholia: <https://scholia.toolforge.org/> [↵](#)
10. The question of how many items about publications Wikidata can include within the limits of the software and whether a separate database would improve performance is a complex question that is still being debated in the community; a separate database would increase capacity, but lead to additional questions of interoperability and maintenance. See for background of the issues the WikiCite roadmap : <https://www.wikidata.org/wiki/Wikidata:WikiCite/Roadmap> ; for current discussion a ticket in the Wikimedia technical community on whether items about scholarly articles should be split from Wikidata: <https://phabricator.wikimedia.org/T281854>; and for an outline of one potential future, the Shared Citations proposal, which gathered extensive community support: https://meta.wikimedia.org/wiki/WikiCite/Shared_Citations [↵](#)
11. See for instance the Citation Typing Ontology for a discussion of the types of ways citations can be used: Shotton, D. (2010). CiTO, the Citation Typing Ontology. Journal of Biomedical Semantics 1, S6. <https://doi.org/10.1186/2041-1480-1-S1-S6>. [↵](#)

12. Singh, H., West, R., Colavizza, G. (2021). Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia. Quantitative Science Studies 2 (1): 1–19. doi: https://doi.org/10.1162/qss_a_00105 ↵
13. Ederporto. (June 4, 2021). Brazilian Laws: Modeling the Brazilian legislation in Wikidata. Diff: the Wikimedia blog. <https://diff.wikimedia.org/2021/06/04/brazilian-laws-modeling-the-brazilian-legislation-in-wikidata/> ↵
14. WikiCite grant: Balinese WikiLontar. (2020, December 10). https://meta.wikimedia.org/w/index.php?title=Wikicite/grant/Balinese_WikiLontar&oldid=20779287 ↵
15. WikiProject Source Metadata (https://www.wikidata.org/wiki/Wikidata:WikiProject_Source_Metadata) provides a home for current bibliographic models, tools and other in-progress projects on Wikidata, as well as links to the community's mailing list. ↵
16. For software citation principles from the research community, see Smith A.M., Katz D.S., Niemeyer K.E., FORCE11 Software Citation Working Group. (2016). Software Citation Principles. PeerJ Computer Science 2:e86. DOI: [10.7717/peerj-cs.86](https://doi.org/10.7717/peerj-cs.86); for GitHub's announcement, see Nat Friedman. (2021, July 27). <https://twitter.com/natfriedman/status/1420122675813441540> ↵
17. IFLA WikiCite discussion series . Scholarly Profiles, Wikidata and Academic Libraries with Meg Wacha. <https://youtu.be/Y2nETuZlaSk?t=507> ↵