{} wikicite

Citations for the sum of all human knowledge

P. Ayers, D. Mietchen, J. Orlowitz, M. Proffitt, S. R. Rodlund, E. Seiver, D. Taraborelli, B. Vershbow. 2019. *WikiCite 2018-2019: Citations for the sum of all human knowledge*

https://doi.org/10.6084/m9.figshare.8947451

WikiCite 2018 Participants, By Satdeep Gill, CC BY-SA 4.0

About

WikiCite is an initiative aiming to build **a comprehensive knowledge base of sources**, to serve the sum of all human knowledge. This report examines the impact, key milestones, and reach the WikiCite community has achieved over the course of the past year.

The Alfred P. Sloan Foundation generously supported WikiCite 2018 and WikiCite activities through 2020.

Past WikiCite events were supported by the Wikimedia Foundation, Wikimedia Germany and Wikimedia Austria as well as Crossref, the Gordon and Betty Moore Foundation, the Alfred P. Sloan Foundation, and the Science Sandbox, a Simons Foundation initiative.

Introduction

Wikipedia is the world's largest, most widely used online encyclopedia. It is free and open, a vast body of knowledge anyone can contribute to. Wikipedia relies on policies that put a premium on verifiability of the information it holds, a commitment to citations, fact-checking, and accuracy.

How does the Wikimedia movement empower individuals to assess reliable sources and arm them with quality information so they can make decisions based on facts? How do we identify bias or distortions in the application of these verifiability policies? These questions are relevant not only to Wikipedia users but to consumers of media around the globe.

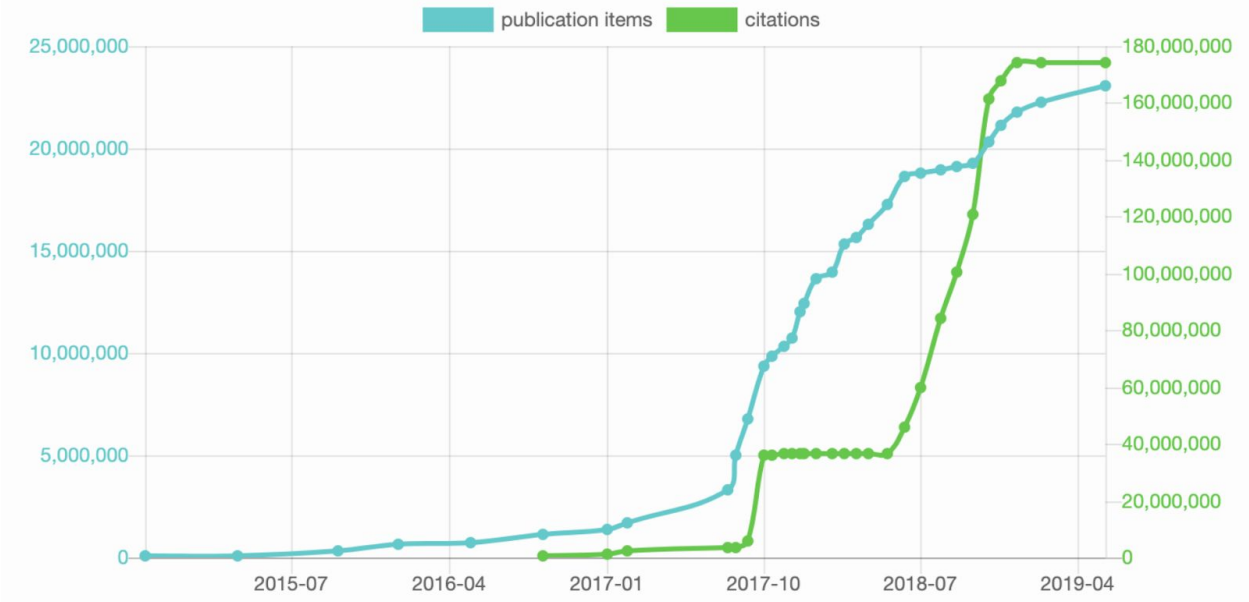
Over the past decade, the Wikimedia movement has come together to answer that question. Efforts to design better ways to support sourcing have begun to coalesce around Wikipedia's sibling project Wikidata – the free knowledge base that anyone can edit. With the creation of a rich, human-curated, and machine-readable knowledge base of sources in Wikidata, the WikiCite initiative is crowdsourcing the process of vetting information and its provenance.

The number of items in Wikidata representing sources has grown dramatically since the first WikiCite conference in 2016, with over 20 million items about publications in Wikidata as of the end of 2018. With this growth has come a large number of innovative tools to work with this data. These tools and experiments have been developed by a community of volunteer Wikimedia developers, librarians, and linked data experts and enthusiasts, and together show the power and potential for this new source of open bibliographic data; these WikiCite-related projects have been presented and discussed all over the world. Additionally, in 2018 there was increasing interest by the library community in using Wikidata and Wikibase (the underlying software of Wikidata) in library-related linked data systems, with several experimental projects being developed. The WikiCite community uses an active mailing list (with 270 members) and social media platforms (with over 3400 followers of the @wikicite account on Twitter), as well as discussions on Wikidata itself.

In November 2018 WikiCite 2018, an international conference was held in Berkeley, California with participants including Wikimedians, librarians, LODLAM practitioners, software

engineers, data scientists, and open knowledge advocates, gathered together the emergent community of WikiCite. This event included a conference day of tracks, a hands-on strategy development summit, and a hack day where participants developed new and existing projects. The WikiCite initiative is unique in bringing together the Wikidata, linked data, and library community together to collaborate on the future of open bibliographic data.

Growth of bibliographic items in Wikidata



From http://wikicite.org/statistics.html as of July 2019

WikiCite 2018

Banner for WikiCite 2018, the annual WikiCite conference • Dario Taraborelli, CC BY-SA 4.0

WikiCite 2018 was a 3-day conference, which included formal presentations, group discussion and a hack day/do-a-thon. The conference took place in Berkeley, California, USA on November 27-29, 2018.

WikiCite 2018's expanded efforts that started with WikiCite 2016 and WikiCite 2017 to explore the vision of creating an open repository of bibliographic data to support the citation and fact-checking needs of Wikimedia projects, and possibly, to serve as an open infrastructure for research, education, and information quality across the web.

2018 presented a pivotal moment for WikiCite, wherein its emergent community grappled with possible growth scenarios, addressed critical gaps, and discussed possible courses for the program's future.

Key figures about the event

- 115 attendees from 18 countries
- 260 applications
- $40K given in scholarships
- 3 full-day strategy tracks
- 24 do-a-thon day demos
- 18 presentations
- 27 lightning talks

WikiCite stickers • Dario Taraborelli, CC0

Participants and organizations

Participants were invited to apply to WikiCite 2018 through an open call. Final selection decisions were made based on a number of criteria, such as applicants' relevant expertise in relation to WikiCite, technical skills, previous contributions to Wikimedia projects, funding requests, gender, demographic distribution, and their personal commitment to fostering diversity and inclusion.

2018 was the first year that the WikiCite annual event was held in North America. To promote the robust participation of individuals from around the globe, we offered more scholarships — with larger amounts awarded to folks traveling from further away — than ever before in the history of the event. The bulk of scholarship funds went to individuals who traveled from outside of the United States to attend.

This year, the organizing committee worked to improve the gender balance at the event. While the majority of applicants were men, we admitted 59% of self-identified female and non-binary applicants and 38% of self-identified male applicants. About 57% of WikiCite 2018 participants self-identified as men, and 43% self-identified as women or non-binary.

Institutional presence at WikiCite included the Alfred P. Sloan Foundation, Chan Zuckerberg Initiative, ContentMine, Crossref, Directory of Open Access Journals [DOAJ], Google, Hatnote, Hypothesis, Internet Archive, ITHAKA / JSTOR, Library.Link Network, OCLC Research, Open Knowledge International (okfn.org), The Alan Turing Institute, and The Open Data Institute.

Academic research and libraries were represented with delegates from ABES (France), American Physical Society, Berkeley Institute for Data Science, California Digital Library, CONICET (Argentina), Cornell University, CUNY, Drexel University, Fondazione Bruno Kessler (Italy), Harvard Library, Indiana University, Los Alamos National Laboratory, MIT, National Library of Wales (United Kingdom), New York University, Smithsonian Libraries, Stanford University, Technical University of Denmark, Télécom ParisTech (France), TIB Hannover (Germany), UC Davis, Uniformed Services University, University of Bologna (Italy), University of Colorado Boulder, University of Illinois at Urbana-Champaign, University of New Mexico, University of Oviedo (Spain), University of Pennsylvania, University of Tsukuba (Japan),

University of Victoria (Canada), University of Virginia, US Holocaust Museum, and ZB MED (Germany).

Several **Wikimedia movement affiliates** participated in the event, with delegates from the Wikimedia Foundation, Wikimedia Denmark, Wikimedia Deutschland, Wikimedia India, Wikimedia DC, Wikimedia NYC, Wikimedia Uruguay, Telugu Wikipedia, WikiProject Med Foundation and Punjabi Wikimedians.

WMF Executive Director Katherine Maher presenting at WikiCite 2018 • Uncommon Fritillary, CC BY-SA 2.0

The Program

WikiCite 2018 was a 3-day event. Each day had a distinct theme:

Day 1 - Conference

Day 1 focused on presenting past and present work on initiatives involving citations and bibliographic data across Wikimedia projects. The conference convened at the David Brower Center in Berkeley, California, USA. Participants from around the globe gathered to attend presentations and participate in lightning talks centered around three key facets of the WikiCite movement: Corpora and databases, tools, and content gaps. The day started with an opening review by Dario Taraborelli, who was then Head of Research at the Wikimedia Foundation, and culminated with a keynote by Wikimedia Foundation Executive Director Katherine Maher.

Day 2 - Summit

Day 2 focused on discussing future priorities and directions for the WikiCite project. The day opened with a call for participants to *Imagine a Future*. Participants spent the day attending one of three strategy tracks focused on three possible future scenarios presented by the WikiCite organizing committee: Scenario 1 - A database of all sources cited in Wikimedia projects, Scenario 2 - A database for curated bibliographic corpora, and Scenario 3 - The bibliographic commons "moonshot." Participants could alternatively attend a day-long track with technical tutorials.

Day 3 - Hack Day / Do-A-Thon

Day 3 gave participants an opportunity to come together and take action in support of some of the themes and scenarios they discussed throughout the event. At the beginning of the day, idea pitches (some originating in concrete proposals made at the report-outs from the Day 2 strategy tracks) were made for a variety of work sessions that took place throughout the afternoon. At the close of the Do-A-Thon, groups reported back about the work they had done.

Katherine Maher's keynote

In her keynote address, "Wikimedia as the Essential Infrastructure of the Ecosystem of Free Knowledge," Katherine Maher, the Wikimedia Foundation's Executive Director, said WikiCite's progress is closely – and crucially – connected to the long-term progress of the Wikimedia projects. The foundation's long-term goals are articulated in its 2030 strategy, which outlines how, over the coming decade, the Wikimedia projects will become "the essential infrastructure of the ecosystem of free knowledge". Reaching new communities and contributors, including those outside of "structures of power and privilege," is paramount to this strategy – and by leveraging Wikidata and building an open knowledge base of citable sources, WikiCite's community is helping to build a critical component of this greater network of connected knowledge, Maher told attendees at the David Brower Center.

"If you have access to all the information that's on Wikipedia," Maher said, "what you often find is that more than 60 percent of citations are not free to read, which means that Wikipedia's promise as a starting point for learning . . . and for doing deeper research is actually stopped in more than 50 percent of cases by paywalls or toll-access publishing."

Referencing the 2030 strategy as she spoke on stage, Maher said, "We have the strategic pillars of knowledge as a service and knowledge equity, which call on us to build the tools and the community, and to take down the barriers to participation to make Wikipedia, in fact, the essential infrastructure of free knowledge. And an infrastructure that elevates all knowledge. It's not just about the technical infrastructure but about the community and social infrastructure that makes this possible. . . . . The community that you're a part of, WikiCite, and the broader Wikidata network, is vital to fulfilling this vision of this world where we all have access, and *participatory* access, to knowledge."

Lightning talks

Throughout the duration of the event, organizers hosted a series of **lightning talk sessions**, to let participants give short presentations of their work, pitch new ideas, or report on progress made during the event.

Social Events

Participants were also invited to meet and socialize formally and informally. On Day 2, we organized dinner and drinks at a nearby venue for attendees to meet and greet each other.

The **complete program** for WikiCite 2018 can be found here.

Friendly space policy

The event was governed by the Wikimedia Foundation's Friendly Space Policy, a policy dedicated to "providing a harassment-free venue and conference experience for everyone, regardless of gender, sexual orientation, gender identity or expression, disability, physical appearance, age, race, ethnicity, political affiliation, national origin, or religion — and not limited to these aspects." No violation of this policy was reported to the organizing committee during or after the event.

Scenario discussions: summary and outcomes

A feature of the WikiCite meeting was the Summit where attendees were able to spend an entire day collaboratively exploring three different potential endpoints for the WikiCite efforts.

- Scenario 1: A database of all sources cited in Wikimedia projects
- Scenario 2: A database for curated bibliographic corpora
- Scenario 3: The bibliographic commons moonshot

Full documentation for the day's discussions are in the linked Etherpads above. What follows is a brief summary of those discussions and some overall observations for potential fruitful project directions.

Scenario 1: A database of all sources cited in Wikimedia projects

This is the only scenario that is explicitly tied to current Wikimedia projects, and where discussions were tightly correlated with an assumed reliance on Wikidata. Here *key opportunities* were seen in moving citations from strings to Wikidata identifiers, which could

facilitate research (and corresponding action) on the citation base that forms Wikipedia. The discussion also revealed opportunities for tool development to facilitate the creation of citations / reuse of existing Wikidata identifiers. *Perceived risks* in this scenario are that tightly coupling Wikipedia with Wikidata has the potential to irritate or disrupt both the Wikipedia[1] and Wikidata user communities. There are also likely to be performance issues in terms of Wikipedia pages with lots of references getting their reference metadata served from Wikidata, but these have not been explored in detail.

Scenario 2: A database for curated bibliographic corpora

Unlike Scenario 1, Scenario 2 referred to Wikimedia projects in only a glancing way. In this discussion, Wikidata and Wikibase were referred to almost interchangeably. *Key opportunities* included collaboration with motivated data curators, and the creation of large corpora of linked data that would facilitate research that would be of broad interest. *Perceived risks* are around lack of ownership clarity over the corpus: is this Wiki driven, curator driven — who is actually driving this train? Wiki owned, curator owned, or some mix where no one has ownership? A stumbling block for discussion participants was arriving at a commonly understood definition for corpora.

Scenario 3: The bibliographic commons moonshot

This scenario captured a lot of attention and was attended by more conference participants than the other two. Discussion notes reveal scant mention of Wikipedia or other Wikimedia projects, apart from Wikibase. This scenario was described by participants as a vision of "one datasource to rule them all." *Key opportunities* for research and broad utility abound with this scenario, as does the tantalizing prospect of doing away with paywalls for metadata. This scenario was seen as a "big tent" that would draw many of invested and motivated participants. As with Scenario two, *perceived risk* includes issues around management. Additionally, discussants saw a risk that commercial players would be able to capitalize or "enclose" this treasure trove of data. Clarity around what is in or out of the "bibliographic commons" seemed murky. It is clear that reaching an endpoint of bibliographic nirvana will require a massive tool effort.

[1] Most especially the English language and other larger language Wikipedia communities. There is a strong possibility that this would be very welcome within smaller language Wikipedia editing communities.

Overall observations flowing out of scenario discussions

The three scenarios covered in the Summit represent three different potential paths for the WikiCite effort. To date, WikiCite has benefitted from a broad vision, but a more focussed approach may yield some more concrete results. The next phase should help to develop a clear end to the sentence, “WikiCite is…” (or even “WikiCite is not…”). In order to develop clarity around this, some structured exploration around stakeholders in different scenarios may be useful in order to identify which communities stand not only to benefit but what valuable opportunities are available through this effort, or to identify where communities can stop investing individually and instead co-invest.

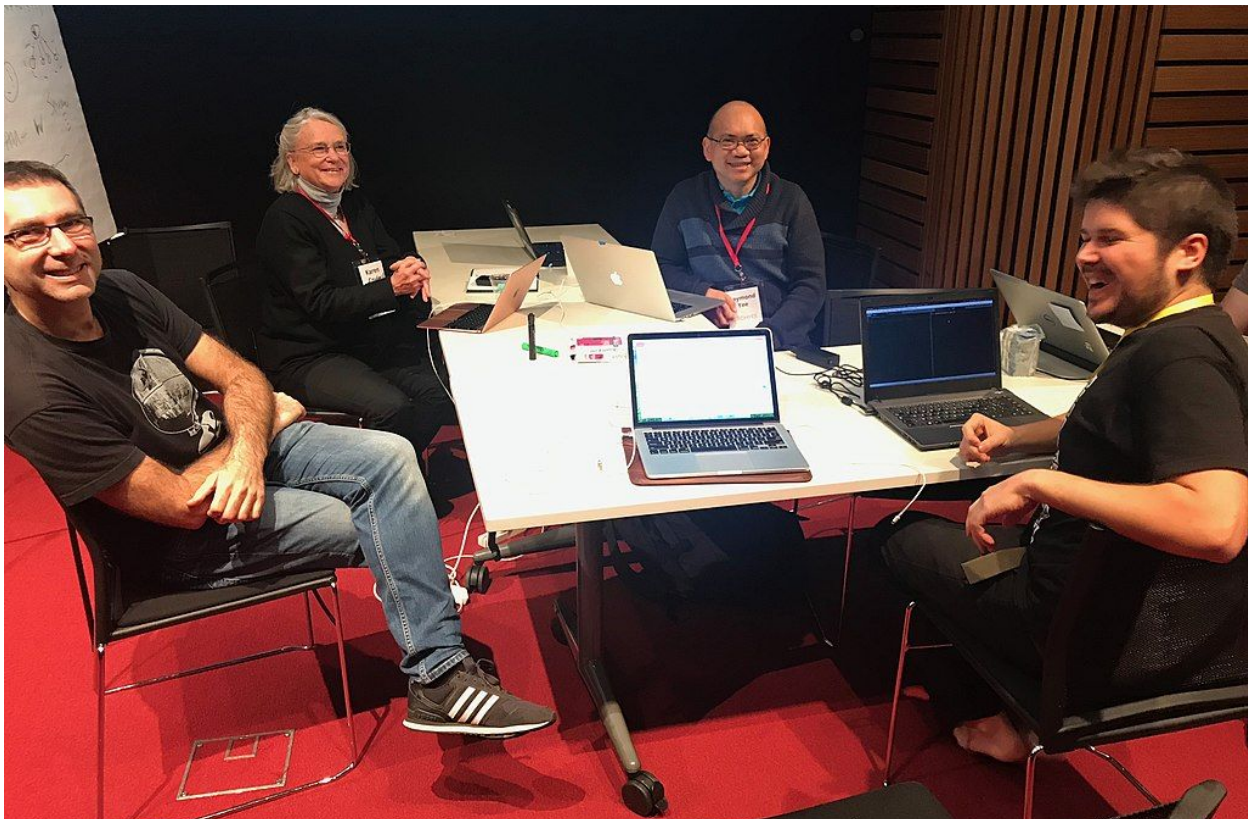
Discussions would be aided by clear definitions for the terms “bibliographic commons” and “corpora.” Several of the conversations discussed how to model citation data in Wikidata (or Wikibase) and some focussed effort could fruitfully be devoted to this goal. Similarly, data synchronization was discussed in all three scenarios as something that would need to be solved in order to move forward.

Preservation and dissemination

Ahead of the event, we created a structure of linked notepads, which were collaboratively edited in real time and represent a comprehensive record of each session. This allowed us to document the full life cycle of the event.

In addition to written notes, all of the main sessions (with the exception of the breakout groups and the work day sessions) were recorded and livestreamed by Andrew Lih. All videos of the event are available under a Creative Commons Attribution license.

Links to event documentation and videos can be found through the WikiCite 2018 program page.

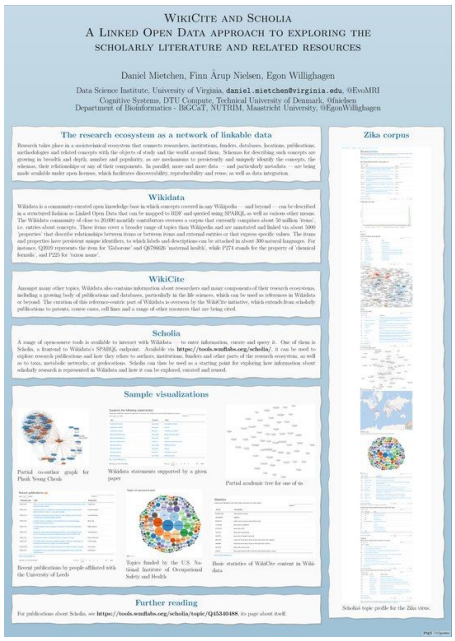


Workgroup at WikiCite 2018 in Berkeley, CA • Dario Taraborelli, CC0

News from the movement

The WikiCite annual conference is only a small part of the story. Over the past year, the WikiCite movement, our partners, and our collaborators spearheaded a number of important initiatives, including: new and enhanced tools for curating structured data on sources, outreach and training events, and pilots for institutional adoption.

Scholia


WIKICITE AND SCHOLIA
A LINKED OPEN DATA APPROACH TO EXPLORING THE SCHOLARLY LITERATURE AND RELATED RESOURCES

Daniel Mietchen, Finn Årup Nielsen, Egon Willighagen

Data Science Institute, University of Virginia, daniel.mietchen@virginia.edu, @EvoMRI
Cognitive Systems, DTU Compute, Technical University of Denmark, @finnlezen
Department of Bioinformatics – BiGCaT, NUTRIM, Maastricht University, @EgonWillighagen

The research ecosystem as a network of linkable data

Wikidata

WikiCite

Scholia

Sample visualizations

Further reading

Zika corpus

Scholia is a tool to explore and visualize scholarly and bibliographic information in Wikidata. The Scholia Web service creates on-the-fly scholarly profiles, e.g. for institutions, topics, events, research projects, and academic journals or individual articles. To collect the data, it queries the SPARQL-based Wikidata Query Service. Among several display formats available in Scholia are lists of publications for individual researchers and organizations, publications per year, employment timelines, as well as co-author and topic networks and citation graphs. The Python package implementing the Web service is also able to format Wikidata bibliographic entries for use in LaTeX/BIBTeX.[2]

Across several Wikipedias, variants of Template:Scholia point to Scholia profiles for the subjects of Wikipedia articles. At WikiCite 2018, Scholia was enhanced with functionality to visualize clinical trials.

[2] Nielsen, FA, Mietchen, D., Willighagen, E., 2017. Scholia and Scientometrics with Wikidata. *Zenodo*. https://doi.org/10.5281/zenodo.1036595

Scholia was also used in the context of pilots to explore the possibility of generating structured data profiles for entire academic schools or institutions[3] or to improve the visibility of underrepresented scholars and their scholarly output[4].

In 2019, the Scholia team received additional funding from the Alfred P. Sloan Foundation to "robustify Scholia" with back-end development and testing based on pilot corpora. The main objective at this stage is to attain stability in challenging cases such as the handling of large or incomplete datasets, or server throttling due to high frequency of queries. Further goals include integrating Scholia with data curation and manuscript writing workflows, serving more languages, generating usage stats, and documentation of workflows.

Scholia has inspired the development of other Wikidata frontends that apply the Scholia approach of embedding the results of predefined SPARQL queries to other domains. For example, Ordia visualizes information around lexemes and their forms, senses and languages in this way.

Wikidata and Wikibase as library infrastructure

One of the most important developments, arising in parallel with the WikiCite effort, is the "rise of Wikidata" as an emerging source of Linked Open Data for libraries and library organizations.[5] We have seen a rise in interest in Wikidata and local Wikibase instances for managing bibliographic data, particularly for managing names and other authority metadata. The forthcoming report *Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage* summarizes several examples, including a local Wikibase instance as a technical backbone the digital humanities project Enslaved.org, and the FactGrid project at the Universität Erfurt, which uses a local Wikibase to record factual data for historians.

[3] Lemus-Rojas, M. and Odell, J.D., 2018. Creating Structured Linked Data to Generate Scholarly Profiles: A Pilot Project using Wikidata and Scholia. *Journal of Librarianship and Scholarly Communication*, 6(1), p.eP2272. DOI: https://doi.org/10.7710/2162-3309.2272

[4] Lemus-Rojas, M., 2019. Enhancing the Representation of Latin American Women Scholars and their Scholarship in Wikidata. *2019 Seminar on the Acquisition of Latin American Library Materials (SALALM)* http://hdl.handle.net/1805/19749

[5] Smith-Yoshimura, Karen. 2018. Analysis of 2018 International Linked Data Survey for Implementers. Code4Lib Journal, Issue 42, 2018-11-08 https://journal.code4lib.org/articles/13867

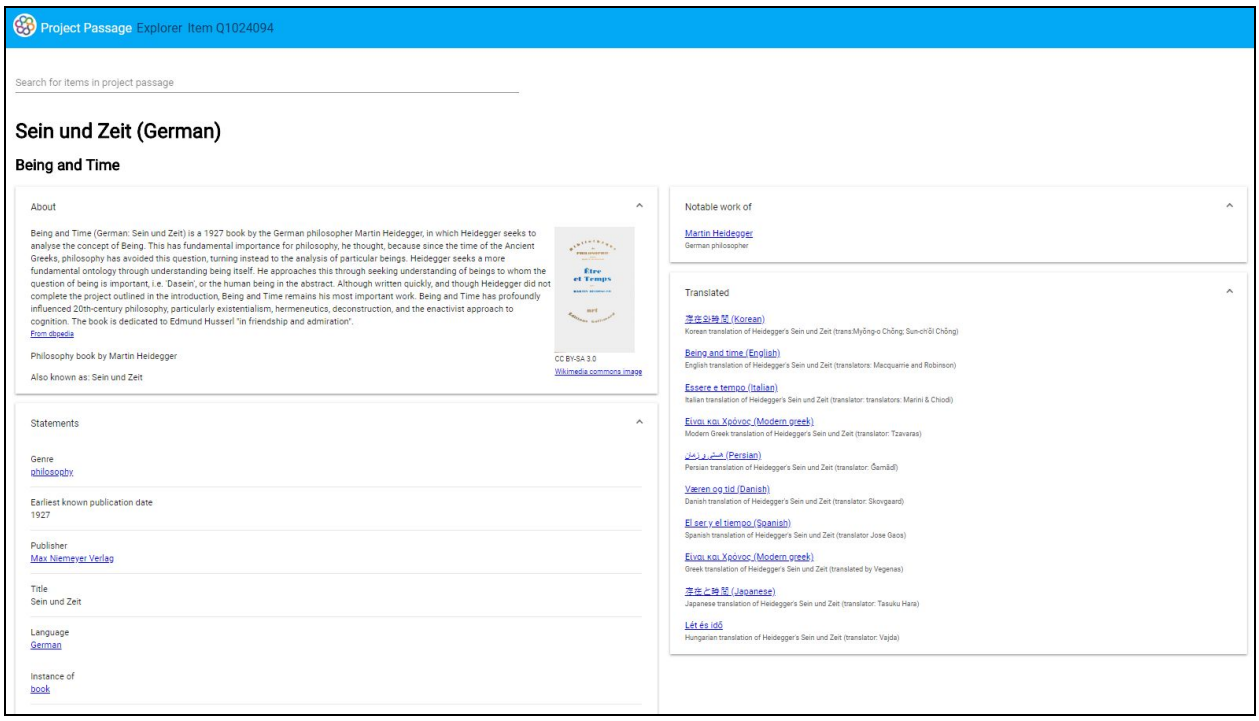
National level-implementations of Wikibase are being developed by the French consortium ABES and the Bibliothèque nationale de France, and by the Deutsche Nationalbibliothek.[6] [7] In May 2019, the Library of Congress announced that it had added Wikidata IDs to *id.loc.gov* authority pages and to the data itself.

Benjamin Bober presents the BNF's Wikibase pilot for bibliographic entities • Dario Taraborelli, CC0

[6] Godby, Jean, Karen Smith-Yoshimura, Bruce Washburn, Kalan Knudson Davis, Karen Detling, Christine Fernsebner Eslao, Steven Folsom, Xiaoli Li, Marc McGee, Karen Miller, Honor Moody, Craig Thomas, and Holly Tomren. 2019. *Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage*. Dublin, OH: OCLC Research. (forthcoming)

[7] Barbara Fischer, Jens Ohlig, 2019. New testing ground for Wikibase: A federal agency goes on an expedition in the Wiki universe. *Wikimedia Deutschland Blog.*

In 2017 and 2018, OCLC's Project Passage worked with member libraries to develop a prototype to demonstrate the value of linked data for improving resource-description workflows in libraries.



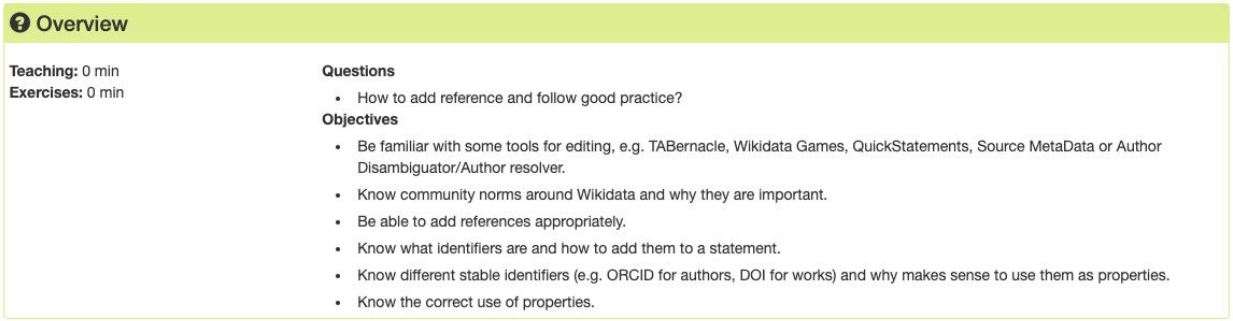
Example from Project Passage: Explorer view of Being and Time with multiple translations

The OCLC team built a prototype service on the Wikibase platform to provide a reconciliation service – to connect legacy bibliographic information to linked data entities and an editor service – to view, create and edit linked data descriptions and relationships. Work is now complete on this initiative; the report *Creating Library Linked Data with Wikibase* summarizes experiences and findings from the project.

In April 2019, the Association of Research Libraries issued a White Paper on Wikidata: Opportunities and Recommendations. The paper, which was written by a group of librarians and Wikimedians, suggested ways that libraries could use Wikidata to engage around collections, collection descriptions, and with faculty.

Wikidata training for librarians: Library Carpentry and LD4

During the strategy session on Day 2 at WikiCite 2018, several participants independently raised the issue of needing to teach more librarians how to use Wikidata and Wikibase. They joined forces for the do-a-thon on Day 3 and sprinted on a set of Library Carpentry lesson plans. This collaboration has persisted since WikiCite and these lessons are still under active development.



Overview

Teaching: 0 min
Exercises: 0 min

Questions

- How to add reference and follow good practice?

Objectives

- Be familiar with some tools for editing, e.g. TABernacle, Wikidata Games, QuickStatements, Source MetaData or Author Disambiguator/Author resolver.
- Know community norms around Wikidata and why they are important.
- Be able to add references appropriately.
- Know what identifiers are and how to add them to a statement.
- Know different stable identifiers (e.g. ORCID for authors, DOI for works) and why makes sense to use them as properties.
- Know the correct use of properties.

The Andrew W. Mellon Foundation funded Linked Data for Production: Pathway to Implementation (LD4P Phase 2) is working to develop "a cloud-based sandbox editing environment in support of an expanded cohort of libraries to create and reuse linked data" and integrate "library metadata and identifiers with the Web through collaboration with Wikidata." The May 2019 LD4 Conference on Linked Data in Libraries featured several talks and panels that underscored the utility of Wikidata and Wikibase for libraries. Participants in the project created a Wikidata Affinity Group aiming to understand how libraries can contribute to and better integrate library metadata with Wikidata.

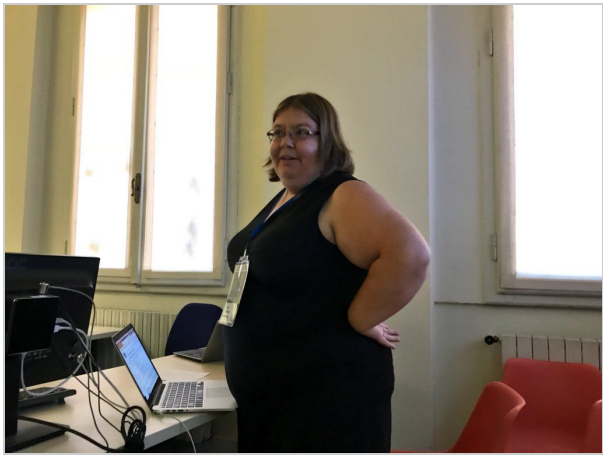
The Wikidata Zika Corpus

The Zika corpus continues to be curated via a dedicated Wikiproject and to grow along with the literature on the topic. It currently comprises about 4,800 articles. On that basis, it continues to be used as an example for testing, learning and demonstration purposes, e.g. to showcase data quality control workflows via ShEx or to explore topic-centric SPARQL queries. In addition to linking publications with their topics, work has intensified on linking publications to their authors, such that no unresolved author name string on Zika papers

currently occurs more than ten times, and queries like institutions of authors who published on the Zika virus before 2016 are yielding more meaningful results. While over 2,300 authors of Zika papers have been identified, there are still about 19,000 author name strings on Zika publications to be resolved at this time.

WikiCite and Open Citations

Several WikiCite community members participated in the 2018 Workshop on Open Citations and hackathon (Bologna, September 3-5, 2018). Dario Taraborelli gave the opening talk on *Remixing the Graph*, discussing opportunities for citation data unlocked by the Initiative for Open Citations to be reused in open knowledge projects like Wikidata. WikiCite-related projects that were worked on at the workshop include: visualizing the citation graph structure (cocitation, coauthorship, bibliographic coupling) of 5000+ papers in the Wikidata Zika Corpus; developing a REST API for bibliographic resources in Wikidata to support analysis via VOSviewer; creating a new project to model and represent retractions in Wikidata.



WikiCite participants at the Workshop for Open Citations 2018 • Dario Taraborelli, CC0

Wikidata-driven bibliographic templates

Some of the participants in the do-a-thon at WikiCite 2018 experimented and improved on the Cite Q template, which pulls relevant citation data from Wikidata into Wikipedia. For an example of Cite Q in action, see the Further Reading section in the article on the Eugene Register-Guard. The Cite Q template helps to demonstrate the principles behind WikiCite.


Further reading [edit]

- Warren C. Price (1976), *The Eugene Register-Guard: A Citizen of Its Community*, Binford & Mort, ISBN 978-0-8323-0271-8, Wikidata Q59292561
- George Stanley Turnbull (1939), *History of Oregon Newspapers*, Binford & Mort, Wikidata Q56862211

Tools

Several tools were designed, extended or adapted to support WikiCite activities. In particular:

- OpenRefine is a general-purpose data wrangling tool that has been increasingly integrated with Wikidata and has been used in a variety of WikiCite contexts, e.g. to add affiliation information to author items based on information available from ORCID.

- Inventaire is a Wikidata frontend that allows people to manage metadata about collections of books being both harvested from and edited on Wikidata. Work on Inventaire has spurred the development of a general-purpose Wikidata Command Line Interface (Wikidata CLI) and of the associated Wikidata Hub, which can, for instance, find the DOI that corresponds to the PubMed identifier for a given publication.

- QuickStatements is a general-purpose editing tool that facilitates batch-editing Wikidata. It is used, for instance, by the Author Disambiguator, a tool that helps identify authors of publications. QuickStatements has seen several major waves of development of its own and has been integrated with EditGroups, a tool to facilitate quality control of batch edits that is frequently part of WikiCite workflows and was in turn developed to facilitate Wikidata editing through OpenRefine. SourceMD and ORCIDator are QuickStatements-based tools to curate Wikidata items about publications and their authors, respectively, based on information available in

external databases like Crossref or ORCID. Both have been considerably modernized over the last year, and their functionality was merged into a unified tool still under the name SourceMD. QuickStatementsBot has similar functionality but is fully automated. Citationgraph_bot 1 & 2 have been the main contributors to the citation graph in Wikidata but they have been inactive since late 2018, much like Fatameh, a tool similar to the older SourceMD that facilitates the curation of Wikidata items about publications.

- Shape Expressions provide a mechanism for quality control across RDF graphs, and their development on Wikidata has received considerable input from the WikiCite community. Wikidata Shape Expressions Inference is a tool for building such Shape Expressions (also known as ShEx) based on a SPARQL query.

Other tools:

- OA Bot enhances references on the English Wikipedia with links to freely available versions.
- Similarly, Cite Unseen marks references on the English Wikipedia as to their nature.
- Reasonator visualizes Wikidata entries within the context of entries linked to and from them, and it links to Scholia profiles for certain kinds of entries.
- Citation.js is a Wikidata-aware JavaScript package for displaying citation information on the web.
- ListeriaBot is a tool to create and maintain Wikidata-based tables on Wikimedia wikis, and it is used, for instance, for prototyping Scholia queries.
- Similarly, TabulistBot creates and maintains tables in tabular data format on Wikimedia Commons.
- Widgets are available to link from the Wikidata UI to some of these tools, e.g. Reasonator and Scholia.

Projects

- Science Stories presents Wikidata-based profiles of women scientists.
- Open Knowledge Maps visualizes information related to a particular topic and has begun to be integrated with Wikidata.
- WikiProject Newspapers collects information about newspapers on both Wikidata and the English Wikipedia.

Events and presentations

Presentations on and about WikiCite were given at many conferences that are either closely allied with the Wikimedia Movement or that are in adjacent spaces, such as SWIB (Semantic Web in Libraries, considered to be the premier conference for linked open data in libraries) and FORCE2018 (FORCE is a preeminent venue for relevant stakeholders engaged around changing how scholarly and scientific information is communicated, shared and used). Below is a non-exhaustive list of WikCite related talks and workshops given during the period of this report.

- WikiData Days 2019, Teresa e Alexandre Soares dos Santos Library, Carcavelos, Portugal 7-8 June. This meeting aimed at exposing intersections in open data and open science projects highlighted WikiCite and Scholia.
- 2019 LD4 Conference on Linked Data in Libraries, Boston, Massachusetts. May 11-12. This meeting featured discussions about the relationship / use of Wikidata and Wikibase in bibliographic infrastructure. See discussion of the LD4 project and conference above.
- Zumstein, Philipp; Voß, Jakob (27 November 2018). "WikiCite". Satellite/breakout session at SWIB18 in Bonn.
- Mietchen, Daniel; Taraborelli, Dario (6 November 2018). "Wikidata and Wikibase as global platforms for democratizing data publishing". — presentation given at the SciDataCon 2018 conference in Gaborone.
- Nielsen, Finn Årup (6 November 2018). "Scholia, WikiCite, Wikidata et al." (PDF). Keynote talk at the Second Linked Open Citation Database Workshop in Mannheim.
- Voß, Jakob (26 October 2018). "WikiCite für Bibliotheken". Introduction to WikiCite for librarians given at Göttingen State and University Library.

- Redi, Miriam; Delpeuch, Antonin (26 October 2018). "Some Fun with Wikicite and WikiCite in Wikidata [TechStorm 2018]". The Hague, Netherlands. Retrieved 2018-10-26. Introduction to WikiCite for the participants at the Wikimedia NL Hackathon, a.k.a the WikiTechStorm
- Taraborelli, Dario; Mietchen, Daniel (12 October 2018). "A hands-on introduction to Wikidata and WikiCite". — presentation given at the FORCE2018 conference in Montreal. This presentation featured Scholia.[8]
- WikiCite Workshop at TechStorm (October 2018), The Hague, Netherlands (TechStorm is a two day Hackathon).
- WikiCite was featured during a September-October 2018 Australian "tour" to cultural heritage and other institutions in Australia: Wikidata Tour Down Under.
- The WikiCite workshop, Wikimania 2018, Cape Town, July 2018 produced the first draft of the 3 strategic scenarios that were further discussed at WikiCite 2018.
- Wikimedia Hackathon 2018, May 2018 Barcelona. The three day Hackathon included an entire track dedicated to WikiCite, including discussions about Scholia, Zika Corpus, tools and other related efforts.

In addition we anticipate that WikiCite will have representation at significant upcoming conferences.

- Wikimania 2019, Stockholm, Sweden
  - Improving discovery of medical journal articles
- WikiConference North America, Boston, Massachusetts
- WikidataCon 2019, Berlin, Germany
  - WikiCite submissions to Wikidata Con
- SWIB 2019, Hamburg, Germany

Future Directions

Opportunities and hopes

[8] Taraborelli, Dario; Mietchen, Daniel (12 October 2018) A hands-on introduction to Wikidata and WikiCite. https://doi.org/10.6084/m9.figshare.7201469

The future is bright for the intersection of linked open bibliographic data and Wikidata. A recent report published by OCLC found that Wikidata was the linked data source with the largest increase in usage over the last three years among 104 linked data projects and services from 23 countries.[9] In 2019, the Association for Research Libraries released a report on Wikidata in libraries[10], and libraries large and small are investigating how to include Wikidata information and IDs in their catalogs and authority files, and how to include authoritative data in Wikidata. Many of these initiatives were developed, explored, and supported at WikiCite.

There is a need for an infrastructure for trusted knowledge: a place that can signal the authenticity of claims and sources, and which is independent, and which can be relied upon, at scale, as a service — like a network of fact-checkers, a clearinghouse of provenance. It may not be too much of a stretch to consider Wikipedia today (however imperfect) as a kind of epistemic backbone of the internet — constantly evolving and entirely driven in its policies and directions by a global community of volunteer contributors.

Moreover, Wikidata is swiftly becoming a critical node in the broader linked open data web. And more than one third of its 50+ million entries represent bibliographic data. This is the possibility that WikiCite first convened three years ago to explore: how Wikidata might serve a broader purpose for tracking the sources of knowledge in the web ecosystem — as open infrastructure for libraries, publishers, scholarly disciplines, and the broader public.

Challenges

The WikiCite organizers have identified a number of challenges that need to be addressed in the future. These include the scalability (both technical and social) of the project on Wikidata, [11] opening new sources of bibliographic data, and linking to existing data sources at libraries and institutions.

[9] Smith-Yoshimura, K. (2018). Analysis of 2018 international linked data survey for implementers. Code4Lib Journal, (42). https://journal.code4lib.org/articles/13867
[10] Judy Ruttenberg (2019) ARL White Paper on Wikidata Highlights Use of Open Knowledge in Scholarly Communication, Special Collections. *ARL News.*
[11] See https://www.wikidata.org/wiki/Wikidata:WikiProject_Limits_of_Wikidata

For individual and institutional contributors, there are many areas to work on to realize the goal of “the sum of all citations”:

- A number of data modeling challenges remain to represent works in Wikidata. Citations are not limited to books and articles, but encompass all kinds of sources, including databases, newspapers, court cases, preprints, and much more. These source types need to be further developed on Wikidata in conjunction with the Wikidata community.
- There are a vast number of sources published in non-English languages and sources without identifiers (such as DOIs or ISBNS) that are not yet represented in Wikidata, and getting open data for these sources (including many sources currently cited on Wikipedia) represents an unsolved challenge.
- As libraries increasingly realize the value of including links to and data from Wikidata in library catalogs and systems,[12] tools will need to be built and outreach will need to be done to include more institutions, particularly libraries beyond major research institutions. Libraries with limited capacity for new metadata projects will need staff training and support in importing and using data from WikiCite.
- We need to continue to engage with stakeholders to expand the initiative with technical partners.
- We must prioritize and serve communities outside North America and Europe who are most in need of collaboratively curated metadata .

Knowledge equity and diversity

WikiCite was seeded with a large cache of knowledge from English Wikipedia about English language subjects, but we want to think about shifting emphasis toward emerging communities. We are now in a strategic position to draw a map of what's currently accessible on the web, who created it, and what it's about. We have an opportunity to see a clear picture of the knowledge we have represented today — but also the knowledge that we don’t.

[12] See, for example, the Library of Congress’s recent announcement of large-scale integration of Wikidata entities in its linked authority data services:

The results display clear gaps in knowledge equity, representation, and participation — and these mirror structural inequities in the world of information at large.

Much like Wikipedia runs on MediaWiki, Wikibase is the open source software that powers Wikidata. And as a local data management tool, Wikidata is built for federation, on open software, using open standards. It lowers the barrier to producing and managing linked open data, by making a small part of a larger system available for localized use.

For the future, we imagine a vast and decentralized network of linked, federated databases each specializing in a type of art, or a region's great music, a catalog of books, or a subfield of scientific research — sharing data back and forth through compatible data models and smart bots, full circle, with humans facilitating the aggregation and accuracy of the pieces that originate from or pass through their hands. This is the ecosystem we need to build, to expand, and to protect.

Achieving this would require the participation of many organizations, some of which need to maintain a certain amount of control over their data, for quality or privacy or other reasons. But Wikibase's inherent opportunity for federation, including the ability to push and pull throughout the network, can be used to encourage open data with local controls and customs. It offers us the chance to connect locally curated knowledge — mapped and shared through Wikidata as a hub of hubs — with many webs weaving together into a true bibliographic commons.

Risks to the commons

An imperative for this community is to ensure that the products of our work — the tools and the data — remain part of an open knowledge ecosystem. As a movement that values free code, free data, free knowledge, and free people, we need to invest in community-created, open collaborative infrastructure. It's critical that we create the tools, partnerships, messaging, and incentives to drive participants to our open ecosystem, rather than proprietary offshoots of it.

If an ecosystem is not "locked open", as Cory Doctorow says, it is at risk of being "locked closed." When core infrastructure and data are tied down under proprietary code and paywalls, membership costs, or restrictive user agreements, the commons loses. There are ways that commercial players, vendors, and corporations can be a supportive part of our

movement, but we need to define and defend our ways of choosing and working with partners.

In the framework of Paul Peters, CEO of open access publisher Hindawi, the infrastructure of the commons needs to be open source, open data, open to integrations, and open in the transparency of its contracts.[13] Similar principles have been recently expressed by the organizations involved in the Invest in Open Infrastructure initiative.[14]

Creating, no less maintaining, this future involves — as Bilder, Lin, and Neylon detail — a multi-pronged effort: running the infrastructure of the commons through community governance, supporting the infrastructure of the commons through sustainable funding, and preserving community ownership of the commons to insure that it remains open.[15]

Plans for 2019-2020

To innovate how WikiCite convenes and engages the community, the organizing committee decided that Year 2 of this WikiCite phase (2019-2020) will focus on a multi-tiered strategy of regional outreach and education, sponsored attendance at related conferences, and activity at satellite events with remote participation. So unlike in past years, Year 2 won't have a sole, central WikiCite convening.

We will use the momentum of last year's conference to spread and expand the WikiCite community, giving users and builders time to invest in their tools and pilots. We will focus on bringing in new voices and regions through targeted trainings. As Year 2 wraps, we will evaluate the successes and challenges of the year's efforts and decide on the best convening strategy for Year 3.

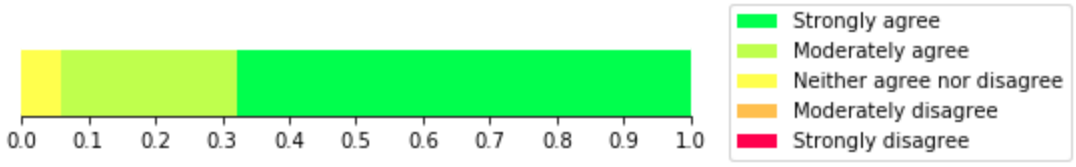
[13] https://about.hindawi.com/blog/a-radically-open-approach-to-developing-infrastructure-for-open-science/
[14] Invest in Open Infrastructure https://investinopen.org/
[15] Bilder G., Lin J., Neylon, C. (2015) Principles for Open Scholarly Infrastructure-v1, https://doi.org/10.6084/m9.figshare.1314859

Appendix 1: WikiCite 2018 Survey

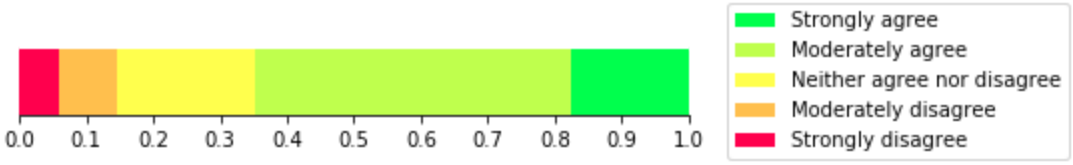
Thirty-six attendees responded to the survey, with 36% of them returning attendees from previous years and 64% there for the first time.

The most common area that people engaged with was defining a technical roadmap and strategic directions. People were overall very satisfied with the conference on Day 1 (80%), and also enjoyed participating in the summit and do-a-thon (44%). The main concern was around timing for the summit and do-a-thon.

I always felt free to voice my comments during the [group activities].



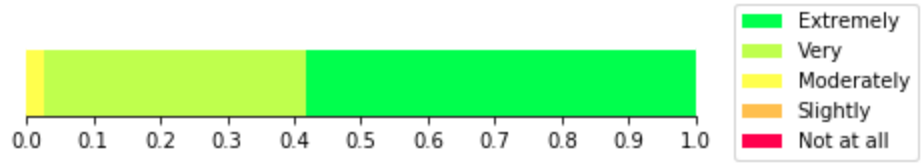
It was feasible to meet my group/session's goals in the time allotted.



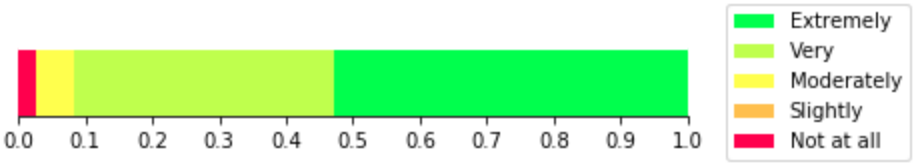
Motivation for participating in WikiCite

Overall, participants were very driven by meeting others and seeing what they were working on.

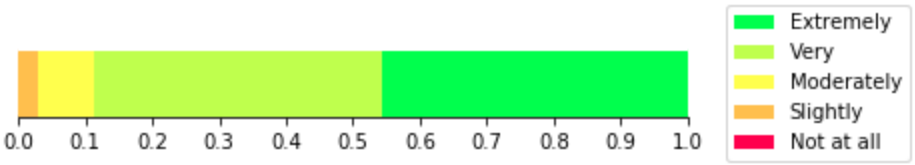
Seeing what others are working on



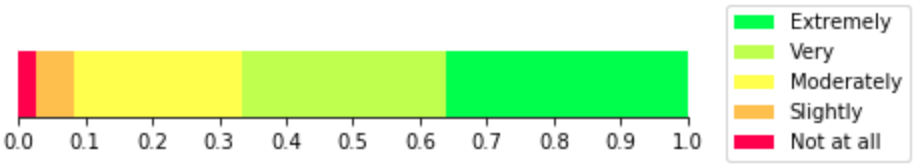
Meeting new people



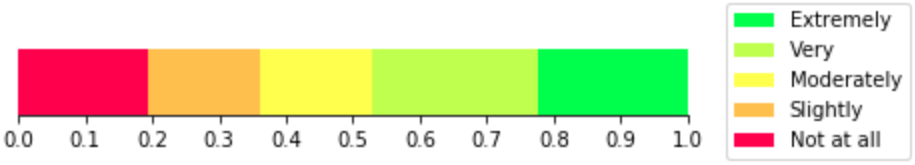
Becoming part of a community



Learning new tools or skills



Dedicated time to get work done



Appendix 2: WikiCite Quotes from participants

What was particularly good about WikiCite 2018?

Participants enjoyed interacting with other enthusiastic attendees, hearing quality talks and learning from speakers, and building community with those from diverse backgrounds.

"It was my first wiki event, and I felt I was already part of the community."

"Meeting and talking with a wide variety of interesting people who had something in common that we were eager to discuss."

They also praised the overall event structure, finding it flexible, inclusive, adaptable but focused, with shared goals in mind.

"[It provided] entry points to all, from experienced community members to newcomers."

"It felt much more productive than typical Wikimedia conferences of a similar size. Most participants were experts of their own domain, a wide range of domains was represented, but at the same time there was a fairly narrow shared goal, so forming collaborations was easy."

What could be improved?

While attendees felt that many different viewpoints and areas of expertise were represented at WikiCite, they wanted more representation of regions outside of the Global North and English speakers. The conference makeup was partly due to entry visa issues for people travelling to the United States, and a need for wider outreach. Attendees also had

constructive suggestions for how to better carry on the goals of WikiCite between meetings and vary event formats.

"I think while this event was great, to duplicate it would be the wrong approach – each event should evolve to reflect the latest state of things as the project itself moves forward."

"Offer clearer templates for doing [something] constructive throughout; focus on one or two things to accomplish (while leaving unconference space for anything else); have follow-up target 6-8 weeks later for publishing proceedings/results."

Financial report

Reporting period: July 1, 2018 - April 30, 2019

Total Sloan Funding	$66,667.00
Total Sloan Monies Spent	$66,667.00
Total Cost of WikiCite 2018	$90,336.83

Cost breakdown

Staff	$16,981.45
Scholarships	$40,586.85
Other	$612.67
Venue-Catering	$32,155.86
Total	$90,336.83

Acknowledgments

WikiCite 2018 was organized by Phoebe Ayers, John Chodacki, Daniel Mietchen, Jake Orlowitz, Merrilee Proffitt, Sarah R. Rodlund, Elizabeth Seiver, Dario Taraborelli, and Ben Vershbow.

The initiative was made possible by generous funding from the **Alfred P. Sloan Foundation**. We are very grateful for their support.

Thank you to Andrew Lih for video documentation, and to our conference venue, the David Brower Center. The organizing committee also wishes to thank the **Wikimedia Foundation** for their support, in particular: Rachel Farrand, Irene Tait, Jonathan Curiel, and Léa Lacroix (from Wikimedia Deutschland).