{} wikicite

Citations for the sum of all human knowledge

D. Taraborelli, L. Pintscher, D. Mietchen, S. Rodlund (2017) *WikiCite 2017 Report*.
doi.org/10.6084/m9.figshare.5648233 • commons.wikimedia.org/wiki/File:WikiCite_2017_report.pdf

WikiCite 2017 group photo • Stephen Laporte, CC BY SA 2.0

About

WikiCite is an initiative aiming to build **a comprehensive knowledge base of sources**, to serve the sum of all human knowledge. In 2017, we convened nearly 100 attendees from 22 countries in Vienna for our annual event, to discuss progress, community needs and technical challenges towards this vision. This report examines the impact, key milestones, and reach the WikiCite community has achieved over the course of the past year.

WikiCite 2017 is generously supported by:

Alfred P. Sloan Foundation

Gordon and Betty Moore Foundation

Science Simons Foundation Sandbox

Wikimedia Foundation

Wikimedia Österreich

Background

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's success lies in the neutrality and verifiability of the information it holds, its rigorous and transparent commitment to citation, 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 in facts? This question is 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 Wikidata – the free knowledge base that anyone can edit. With the creation of a rich, human-curated, and machine-readable knowledge base of sources, the WikiCite initiative is crowdsourcing the process of vetting information and its provenance.

Citation needed • Dario Taraborelli, CC0

The Event

Vienna • Thomas Fabian, CC BY SA 2.0

WikiCite 2017 took place in Vienna, Austria from May 23 - 25, 2017. The event included formal presentations, group discussions and unstructured hack sessions. The event was nearly double in size from the previous year in Berlin, Germany.

Key figures about the event

- 94 attendees from 22 countries
- 39 fellowships
- 16 conference presentations
- 17 workgroups
- 38 lightning talks
- 20 hack day demos

Participants and organizations

We selected participants via an open call for participation. Our goal was to convene around 100 participants who each brought unique expertise and insight to the conference. When selecting prospective attendees, the organizing committee took into account multiple

criteria, such as relevant expertise, technical skills, previous contributions to Wikimedia projects, gender, demographic distribution, and funding requests. We made a concerted effort to increase gender diversity across the spectrum.[1] To increase the geographic and demographic distribution of participants, we secured funding for individuals in need of financial support by issuing 39 fellowships (41% of all attendees).

Beside individual participation, the event saw a **strong institutional presence**, with delegates from over 60 projects and organizations, including: the Collaborative Knowledge Foundation, ContentMine, Crossref, eLife Sciences, Google, the Internet Archive, OCLC, Open Knowledge Maps, OpenCitations, ORCID, dblp, Open Foundation West Africa, Open Knowledge Austria.

A selection of organizations represented at WikiCite 2017

Academic research and libraries were represented with delegates from: the Biodiversity Heritage Library, Drexel University, Federal University of the State of Rio de Janeiro (UNIRIO), the German National Library of Economics (ZBW), Haifa University, Harvard University, Indiana University, University of Chicago, Leeds University, Leibniz University Hannover, Mannheim University, MIT, Politecnico di Milano, Technical University of Denmark, University of Bologna, University of Bonn, University of Illinois at Urbana-Champaign, University of

[1] While the gender ratio of the list of applicants was skewed towards self-identified men, we made an effort to balance gender distribution by offering an acceptance rate of 93% to applicants who didn't self-identify as men, compared to 73% for self-identified men.

Leeds, University of Oxford, University of Tsukuba, University of Würzburg, University of Zaragoza, Yale University.

Several **Wikimedia movement affiliates** participated in the event, with delegates from: Wikimedia Argentina, Wikimedia Australia, Wikimedia Deutschland, Wikimedia DC, Wikimedia España, Wikimedia Foundation, Wikimedia Italia, Wikisource Community User Group.

Participation statistics

166 Applications
130 Invitations extended
94 Participants
39 Scholarships offered (41% of all participants)

WikiCite 2017 participants • Dario Taraborelli, CC0

The Program

WikiCite 2017 was a 3-day event. Each day had a distinct theme:

Day 1 - Present

Day 1 was a structured conference with a series of talks in plenary sessions that aimed to present progress on existing work and initiatives for citations and bibliographic data across Wikimedia projects.

Day 2 - Discuss

Day 2 hosted 17 workgroups, which discussed the technical, social, outreach and policy directions for the movement. The selection and prioritization of topics for workgroups was crowdsourced via a semi-structured process involving voting and clustering by the organizing committee.

Day 3 - Hack

Day 3 was designed as an open hack day, allowing participants to get together and build, based on new ideas and applications

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

We also organized a couple of **social events** at the close of the last two days. On Day 2, we invited two of our funders to take the stage and share their thoughts. On Day 3, we held a closing party where all the participants were invited to socialize and celebrate a successful event.

The **complete program** for WikiCite 2017 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.

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.

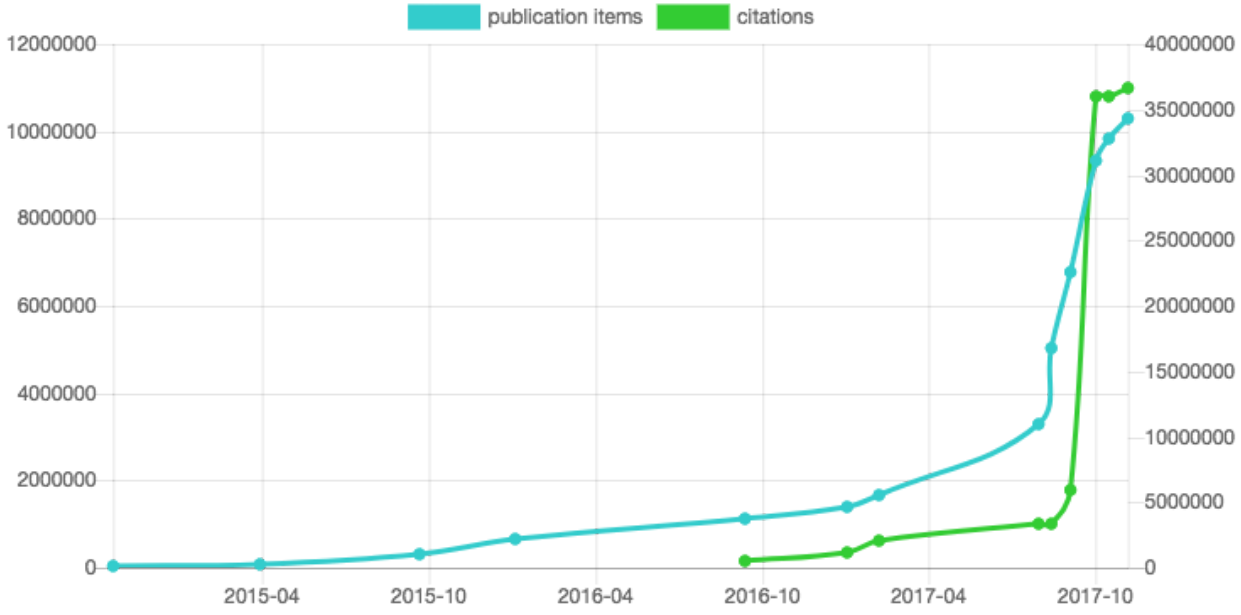
In addition to written notes, all of the main sessions (with the exception of the breakout groups and the hack day sessions) were recorded and livestreamed by Andrew Lih. All videos of the event are available under a Creative Commons Attribution license. A compiled list of these videos, edited by Ewan McAndrew, is available here.

WikiCite 2017 participants • Dario Taraborelli, CC0

Highlights

WikiCite spearheaded a number of important initiatives. This is a list of selected highlights about projects that were either initiated at the event or significantly supported by it.

Growth of bibliographic items in Wikidata



WikiCite statistics

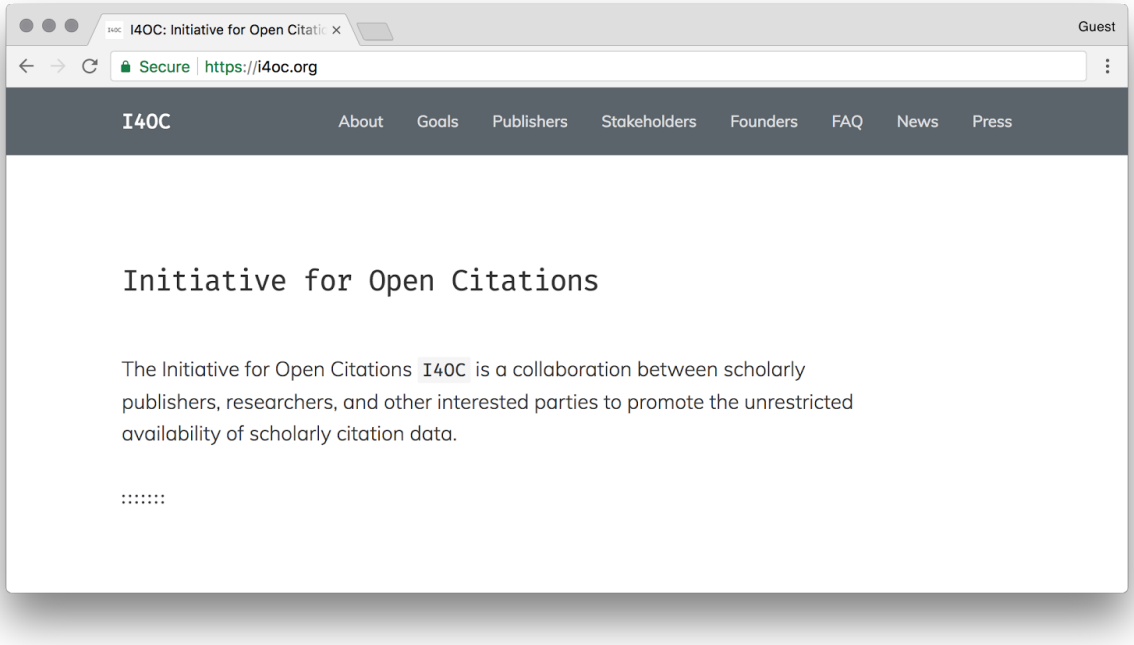
As of November 2017, **1 out of 4 items in Wikidata represents a creative work**. Wikidata currently includes **10 million entries about citable sources**, such as books, scholarly papers, news articles. Wikidata also includes over **1 million author statements** (i.e. statements linking an item about a publication item to the item(s) of its author(s)) and over **40 million author string statements** (i.e. statements linking the item about the publication with a string for the author's name).

Metadata coverage in specific areas competes with canonical databases:

32,840	42,640
Journals indexed in *Web of Science* (source)	Wikidata items about scholarly journals (source)

With a steady growth of items about publications, authors, outlets, we're building a rich knowledge base of sources, linking them up to all the structured knowledge already included in Wikidata, in a way that will allow us to represent where knowledge comes from.

Unlocking references: The Initiative for Open Citations



Initiative for Open Citations

The Initiative for Open Citations (I4OC) was born out of a WikiCite outreach session at COASP 2016 – the annual Conference on Open Access Scholarly Publishing. Launched in April 2017 with the help of members of the WikiCite community, the Initiative for Open Citations made nearly half of all scholarly citations indexed by Crossref – **500 million data points** – openly available to everyone as public domain data. Most of the top 20 scholarly publishers have joined the initiative, which has been featured in *Science Magazine*, *Nature News*, *Wired*, *and Le Monde*, among several other outlets.

We also built a coalition of more than 50 major funders, online platforms, libraries, scholarly and open data organizations supporting the "unrestricted availability of scholarly citation data".

Stakeholders and founders of the Initiative for Open Citations

Data unlocked by the I4OC is actively being reused by several organizations. As of November 2017, 36 million citation links connecting publication items are available in Wikidata. Data from the I4OC is also feeding into **open citation corpora** published by OpenCitations and ScienceOpen.

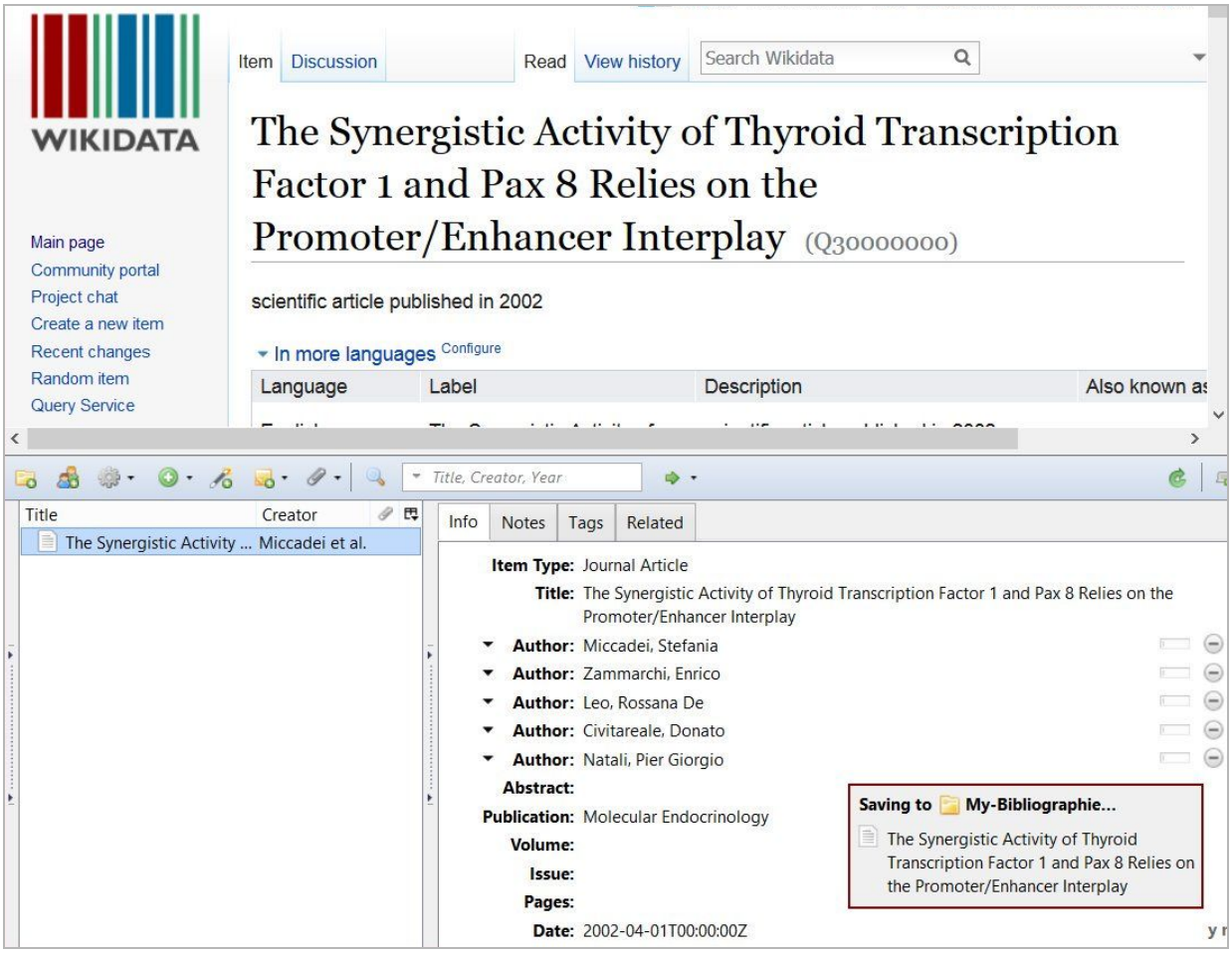
36 million citation links

Wikidata statements representing citations between publication items via the *cites* property [P2860] (source)

Powerful tools to import and reconcile source metadata

A number of tools have been developed, extended or adopted to support a number of workflows involving the creation or curation of Wikidata items about publications, cross-linking and matching of their authors, enriching items with identifiers from external catalogues and authority files, and annotating publications with their topics.

- Creating items about scholarly articles:
  Source MD , Scholia, Fatameh
- Matching paper items with author items:
  ORCIDator, Resolve authors, Source MetaData Game
- Matching paper items with topic items:
  Mix'n Match
- Linking publication items by way of citations:
  Harej's bots
- Adding multilingual information to publication items:
  Mr.Ibrahembot, Emijrpbot, NameGuzzler
- Reporting and monitoring edits on specific classes of items:
  Wikidata SPARQL Recent Changes, Wikidata SPARQL User Edits

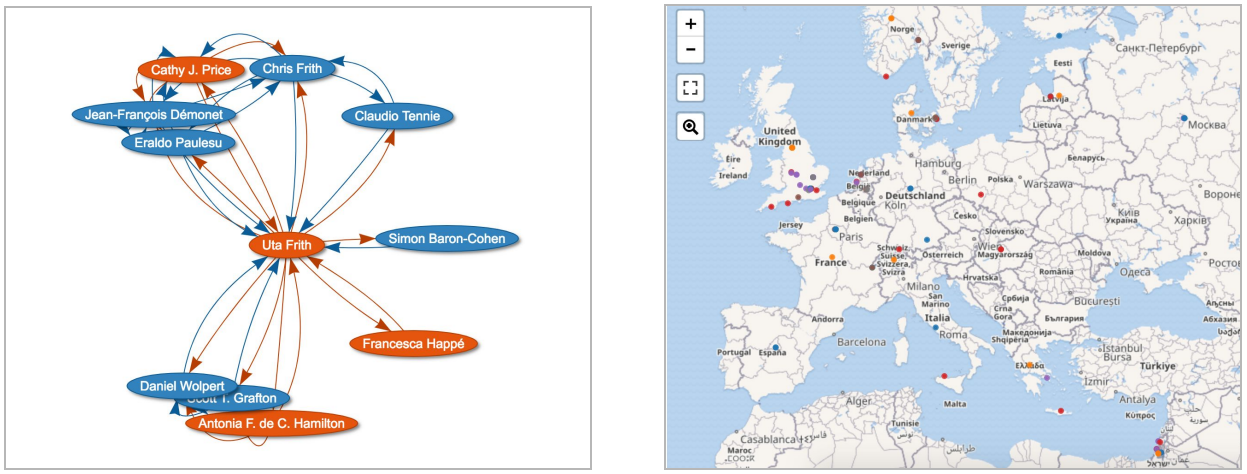


A new translator adding Wikidata support in Zotero • Philipp Zumstein

Open source applications to explore information on sources in Wikidata

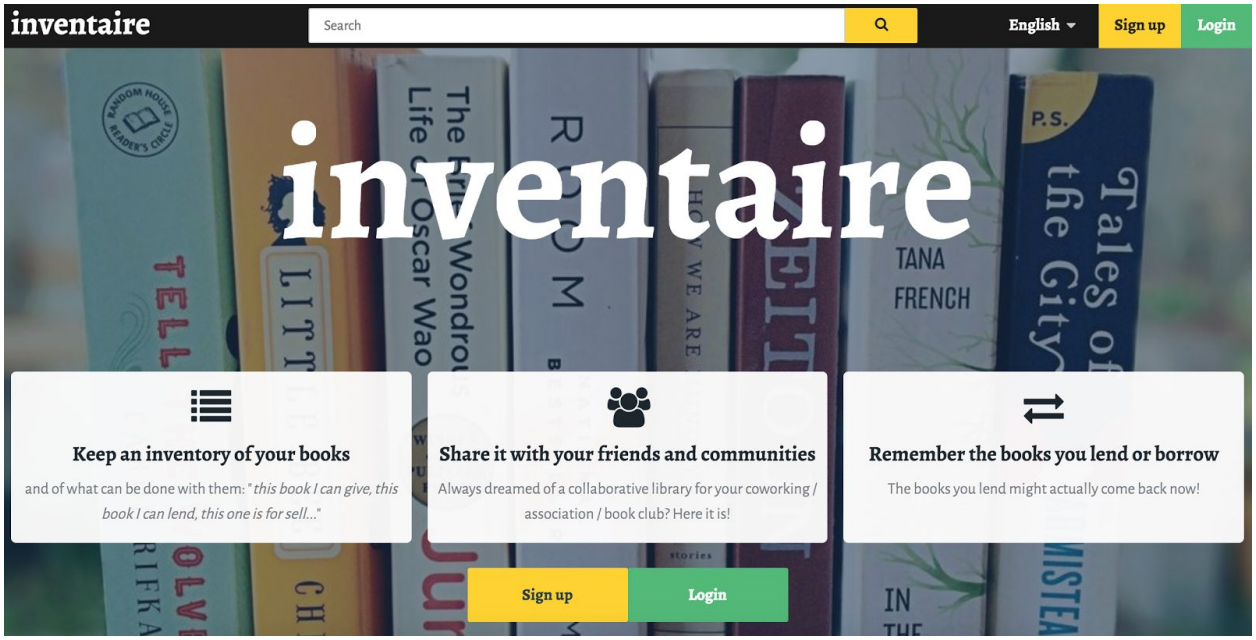
Several applications have been developed to visualize and explore rich bibliographic data available in Wikidata.

Scholia is a collection of interactive tools for exploring and curating bibliographic information in Wikidata. It displays rich publication metadata from Wikidata. It also serves as a reference manager, facilitating proper referencing and quality control, and a metadata import tool. (read more in this conference paper).



Screenshots from Scholia: Uta Frith's co-author graph (left); location of Turing Award recipients (right).

Inventaire is an open source platform for readers to catalog books they own, have read or would like to read, and to share that catalog with their friends and the public in order to facilitate the exchange of such reading materials. The platform builds on open bibliographic metadata from Wikidata, and allows users to contribute to it.



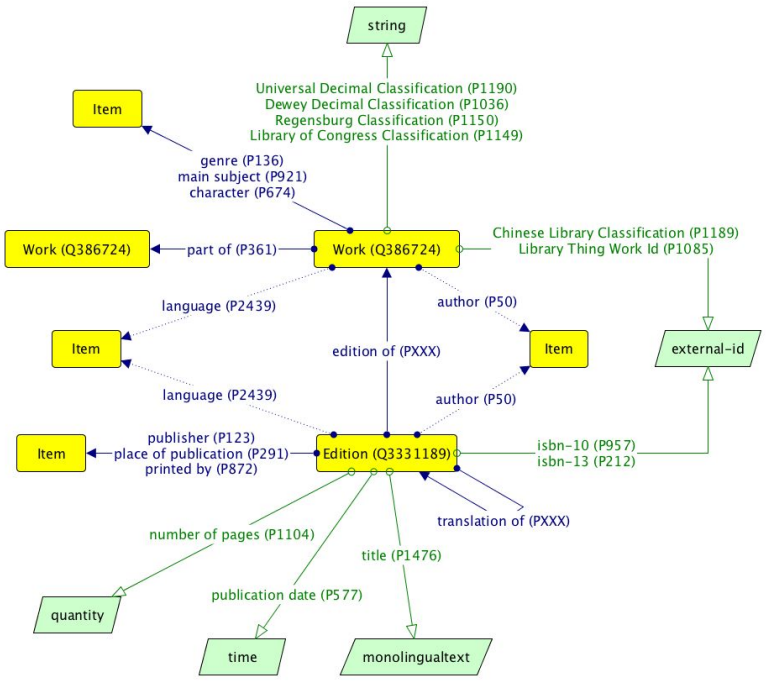
Screenshot from Inventaire

Scholia and Inventaire are great examples of applications enabled by source metadata not subject to copyright restrictions, machine-readable, and openly reusable and editable by anyone.

> "Information on resources should be free and open for citizens to be empowered in their daily choices."
>
> -Inventaire.io

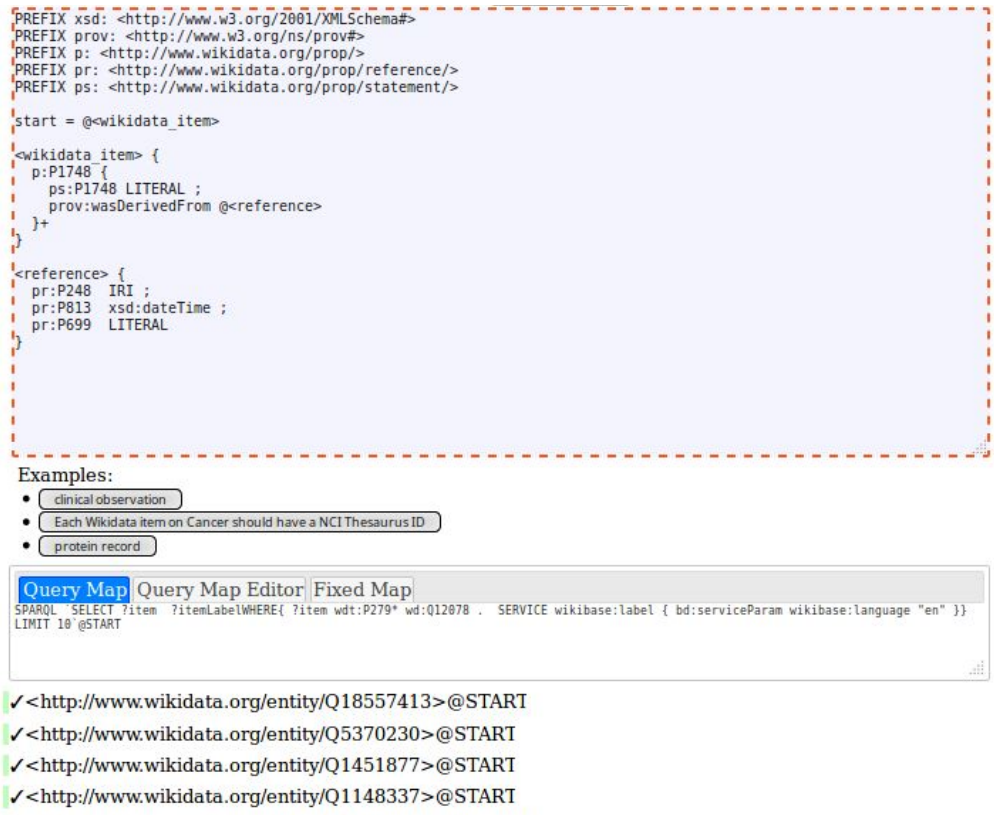
Richer data models and quality control strategies

A working group gathered to discuss a pragmatic **data model to represent books** in Wikidata – a challenge identified since the first WikiCite conference in Berlin. A draft was submitted to the community and is currently under discussion on WikiProject Books. The book data model has been presented at various conferences, including the annual convention of Wikimedians and librarians at the Italian National Central Library in Florence (BNCF) (November 10, 2017), and the Italian Wikicon in Trento (November 18, 2017).



Draft Wikidata book model designed at WikiCite 2017 • Silvio Peroni, CC BY SA 4.0

Another working group explored the possibility of **expressing bibliographic data models and their constraints via Shape Expressions** (ShEx) – a language for describing RDF graph structures. ShEx can be used to validate Wikidata items and communicate expected data-model patterns. It also allows us to collaborate on data modeling, to develop and test models against sample data in agile ways, and to identify inconsistencies or errors. This proposal is now being discussed as a general solution for constraint reports and quality control in Wikidata. The authors of the proposal presented it at a number of conferences including Biocuration 2017 (Stanford, March 26-29, 2017 – slides), Semantics 2017 (Amsterdam, September 11-14, 2017 – slides), and WikidataCon 2017 (Berlin, October 28-29, 2017).



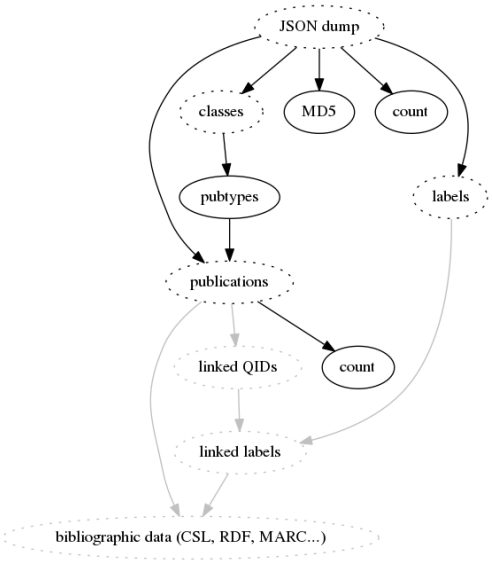
Screenshot of a ShEx validator testing whether Wikidata items returned from a SPARQL query conform to a minimal data model for books. • Andra Waagmeester

Data dumps and statistics

Multiple datasets were created this year to facilitate the analysis and reuse of references and bibliographic metadata from Wikipedia and Wikidata.

A new dataset (Citations with contexts in Wikipedia), released in November 2017, makes structured metadata and contextual information available for every reference added to English Wikipedia articles in a JSON format. Each record represents an individual Wikipedia article revision with all the references parsed, as stored in the English Wikipedia XML dumps. The parsing library to generate the dataset, as well as the JSON schema specification, are also available.

Another project – wikicite-data – first released in August 2017, automates the extraction, transformation, and analysis of bibliographic data from the Wikidata dumps. The project allows Wikidata dumps to be pre-processed and converted to a simplified form, more suitable for use of bibliographic data. It also contains checksums, lists of publication types, and statistics derived from Wikidata dumps.



Data processing flow transforming Wikidata dumps into bibliographic records and summaries • Jakob Voß, CC0

Future Directions

WikiCite is a nascent movement. There is still much work to be done by the community and key partners. The WikiCite organizers have identified a number of challenges that will need to be addressed in the future.

- Different publication types require different data models in Wikidata. As a precondition to populate Wikidata with source metadata, we need the ability to represent every type of work and publication that is cited across Wikimedia projects.

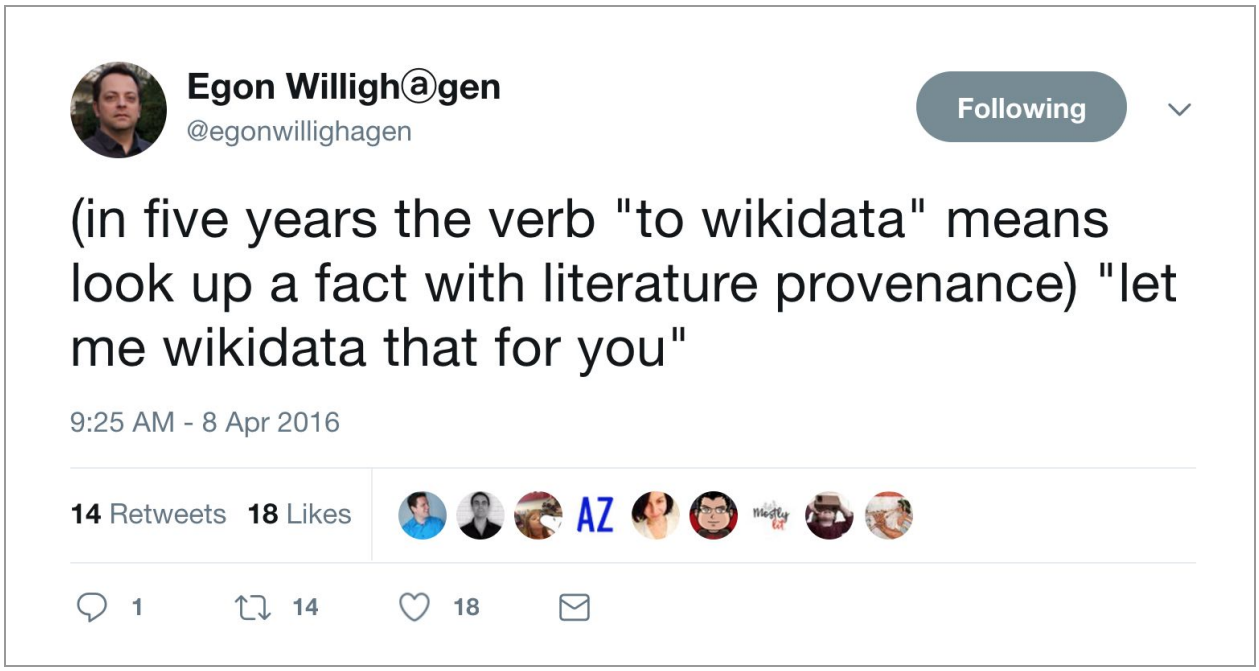
  So far, we have only decided on the modeling for a small number of publication types. To be able to cover more types of sources, and to automate workflows around them, more models need to be suggested, vetted and discussed by the community.

- What tools and processes do we need to import and curate bibliographic information? In order to scale up the work on importing, maintaining and reusing citation data, we need to build additional tools to assist editors.

- How do we get a better understanding of data we already have versus the data we need? To better understand where we should focus our efforts and to make sure we are not introducing harmful biases, we need to better understand the existing citation data in Wikidata. For the citation data to be truly useful, we also need to know which data is needed by Wikipedia and our other data re-users. Then, we can concentrate our work on the most relevant sources.

  As our first goal, we envision a future when every single source cited across Wikimedia projects is captured as a Wikidata item with high quality metadata, identifiers and the appropriate properties.

- We need to better understand how Wikidata and other Wikimedia projects can work together. Different parts of the Wikimedia communities are adopting Wikidata's data at a different rate for many technical and social reasons. We need to focus our efforts on identifying the pain points and fixing them. The reuse of the source metadata in Wikipedia is crucial for the success of WikiCite.



Egon Willighagen on Wikidata (Twitter)

Outreach

Since its launch in 2016, WikiCite and associated projects have been the topic of more than 20 talks and keynotes at national and international conferences, including:

ELAG 2016 • Wikimania 2016 • VIVO 2016 • COASP 2016 • NIH Data Science lecture series • Crossref LIVE 2016 • DLF 2016 • OpenCon 2016 • Convegno delle Stelline 2017 • OER 2017 • Scientometrics 2017 workshop at ESWC 2017 • ElixirCONF 2017 • the 2017 European GLAM-Wiki Coordinators meeting • AAAS 2017 • Biocuration 2017 • Semantics 2017 • Wikimania 2017 • Sfide e alleanze tra Biblioteche e Wikipedia • FORCE 2017 • WikidataCon 2017 • Crossref LIVE 2017 • Biographical Data in a Digital World 2017 • SWAT4LS 2017 • SWIB 2017

Open Science Radio hosted 7 episodes on WikiCite 2017, interviewing participants, organizers, and funders. The event also covered in episode #123 of the Wikipedia Weekly podcast

OSR085 WikiCite 2017 Interview with Carly Strasser [EN]

30. MAI 2017

OSR087 WikiCite 2017 Interview Mark Graham [EN]

6. JUNI 2017

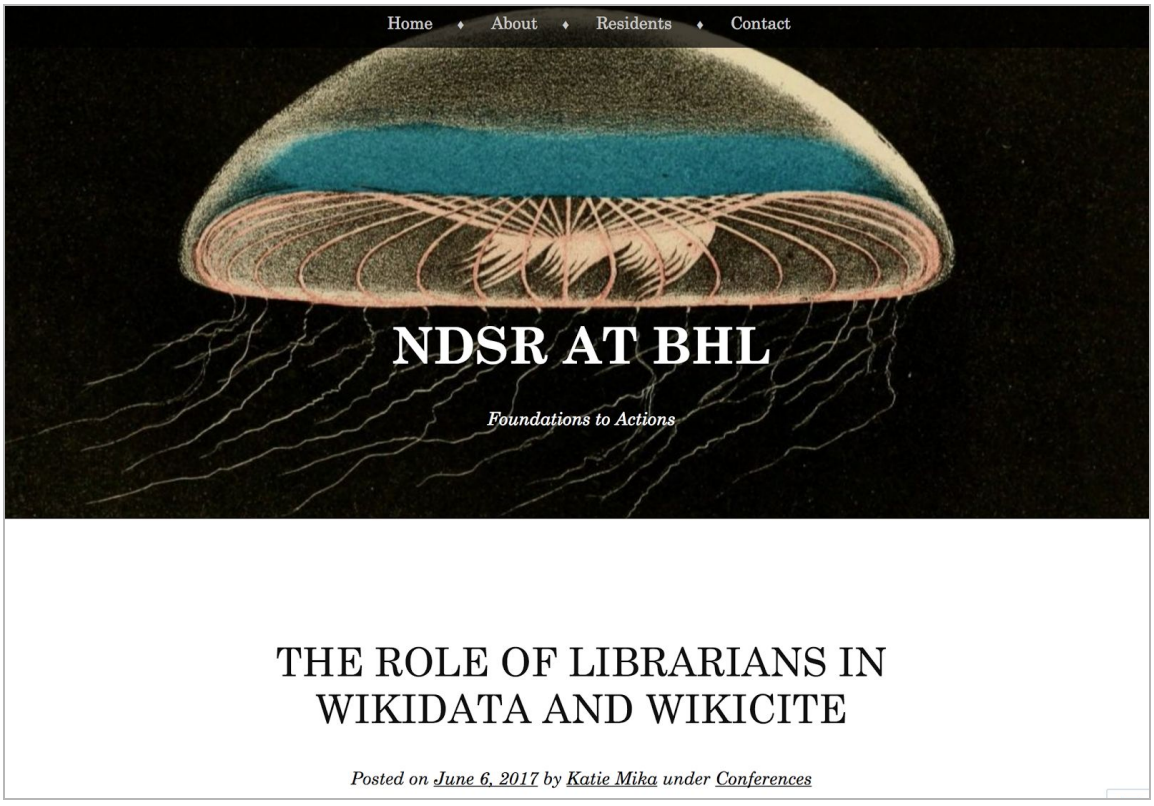
OSR088 WikiCite 2017 Interview Jodi Schneider [EN]

6. JUNI 2017

WikiCite episodes from Open Science radio

Several articles and blog posts from WikiCite 2017 participants were published this year:

- Roderic M. Page (2017) Wikidata, WikiCite, and the "bibliography of life", *iPhylo*
- Philipp Zumstein, Konrad U. Förstner (2017) Tagungsbericht WikiCite 2017. *o-bib*. Bd. 4, Nr. 3 (2017) https://doi.org/10.5282/o-bib/2017H3S106-112
- Finn Årup Nielsen, Daniel Mietchen, Egon Willighagen (2017) Scholia and scientometrics with Wikidata. *Joint Proceedings of the 1st International Workshop on Scientometrics and 1st International Workshop on Enabling Decentralised Scholarly Communication.* https://arxiv.org/abs/1703.04222
- Stephen LaPorte, Mahmoud Hashemi (2017) Wikicite 2017, and the 7 features Wikidata needs most. *Hatnote Blog.*
- Katie Mika (2017) . The Role of Librarians in Wikidata and WikiCite. *National Digital Stewardship Residency (NDSR) Biodiversity Heritage Library*



WikiCite 2017 participant survey

We ran a survey in collaboration with a team of researchers from Carnegie Mellon University[2] to evaluate the event. This survey is is part of a broader project seeking to understand and share best practices for organizing community engagement events, particularly for open volunteer communities. A similar instrument was used to evaluate both WikiCite 2017 and WikiCite 2016.

The 2017 survey received 76 responses in total[3] (80% response rate). We saw continuity in community attendance between 2017 and 2016 events: 35% of respondents to the 2017 survey had also attended WikiCite 2016.

Engagement

There are several considerations for a successful event: overall event organization (such as logistics details), as well as matching the event goals with the expectations of the participations and supporting participants in achieving their ideal outcomes. WikiCite 2017 scored high on all of these dimensions: overall, 98% of respondents were satisfied with the overall event organization, with remaining 2% indicating a neutral response. Additionally, 91% said their expectations going into the event were met, and 87% said their ideal outcome going into the event was achieved. Most respondents said they made at least 2 or 3 new connections with whom they intend to continue collaborating after WikiCite 2017. In particular, consistently more than 85% of respondents said organizers did well in terms of event logistics, and liked the variety and quality of the lighting talks, summit sessions, and Hack-a-thon projects during the event.

Participants' motivation

More than 80% of participants' attendance at Wikicite 2017 was motivated by sharing their experience and expertise with others. Additionally, approximately 60% of participants said

[2] Instrument designed by Anna Filippova, advised by James D. Herbsleb. Supported by Sloan Foundation.
[3] Because all questions were optional, all statistics are presented as proportion of respondents of each question.

they attended to learn new tools/skills, meet new people and become a part of the community. (These options were not exclusive).

Structure of the event

An important aspect in supporting a successful time-bounded event like WikiCite 2017 is encouraging participants to prepare for the event in advance. 63% of respondents said that they read up on relevant literature prior to the event, and 58% said they prepared specific project ideas, questions, or use cases for the event.

Tutorials and community focused themes were a popular aspect of the Summit during Day 2. The best attended Day 2 Breakout sessions according to survey respondents were:

- **Session 1.** How to extend Citoid to improve citations (20% of respondents) and Wikidata 101 (20% of respondents)
- **Session 2.** Introduction to SPARQL and Wikidata query service (30%)
- **Session 3.** Integrating Wikicite into existing activities with wiki-external parties (20%)

For a successful hackathon with a diverse set of attendees, such as Day 3 of WikiCite 2017, it is important to encourage participation and be welcoming of all ideas from all parties. 81% of participants of the Hackathon sessions agreed that the groups were welcoming of all ideas and suggestions, no matter how unconventional. Around 90% said that they felt comfortable to make comments and contribute to discussions. Clarity and feasibility of goals is another important criterion for a successful Hackathon event: about 75% of respondents were clear about the goals of their Hackathon groups and satisfied with the quality of work the group completed. Furthermore, 62% said the timeframe for completing work was feasible.

What were the themes participants engaged the most with across all 3 days of WikiCite 2017? Between 50-60% of participants said that their activity across all 3 days of the event focused on these two areas: 1) defining a technical roadmap/strategic direction for efforts on Wiki citations and source metadata (50% of respondents); and 2) developing tools for citations and source metadata for wiki projects (43% of respondents) (Options not mutually exclusive). These are likely important community themes to follow-up on and consider in future events.

Quotes from participants

Selected feedback from the WikiCite 2017 participant survey

“[I enjoyed] the diversity of the topics covered during the conference and sincerity of the community members.”

“There really AREN'T similar events -- that's what makes WikiCite amazing to me.”

“The breadth of key organisations represented - provided a voice from the many areas with skin in this game”

“I met some brilliant people, learned about a bunch of projects, and made good progress during the hack day. A few of the talks were outstanding, too.”

“[I liked the] depth and range of participants' ideas, skills, backgrounds and expertise. The very high level of commitment, esprit, and inclusiveness.”

WikiCite 2017 participants • Dario Taraborelli, CC0

"The Internet Archive was honored to participate in the WikiCite 2017 gathering where we were able to advance efforts to enhance and support the addition of linked references, with rich metadata, to Wikipedia pages and services. We believe all content that can be linked and referenced should be, and in our perfect world every resource (book, music, software, TV news, academic paper, webpage, etc.) that is available via the Internet Archive and is referenced in a Wikipedia page, should be a click away. Working together we can realize our vision of Universal Access to All Knowledge. Let's turn all citations blue!"

–Mark Graham
Director, The Wayback Machine at the Internet Archive

❧

"At WikiCite I was able to participate in big-picture discussions with colleagues from around the world, as well as do hands-on work to design structured citations. I believe this work will ultimately have impact well beyond Wikipedia and Wikimedia, in making data about scholarly research works and citations more open and available, and in developing the infrastructure and tools to build open citation platforms – something that will open a new frontier for libraries and researchers, who currently rely on a fractured landscape of proprietary, expensive and incomplete tools to find out about publications and research."

–Phoebe Ayers
former Wikimedia Trustee and librarian, MIT Libraries

❧

"WikiCite must become the go-to site for finding scientific publications. Wikimedia search technology is simple and powerful and, in conjunction with the 35 million items in Wikidata, represents the next generation of bibliographic infrastructure. ContentMine is proud to have contributed over 9 million items through WikiFactMine."

–Peter Murray-Rust
Founder, ContentMine

"The idea that many organisations and people hold a little piece of the WikiCite jigsaw puzzle was only strengthened at the event in Vienna. As well as being rather star struck to be sitting at the table with major organisations such as OCLC, Internet Archive, Crossref and the wonderful Wikidata community, I realised this was no simple project.

WikiCite was a fantastic experience with many connections to my professional life as a librarian, educator and researcher - plus an opportunity to engage in person as a member of the Australian chapter. Obviously ensuring Australasian scholarly and trade publishing is included in the WikiCite equation is a priority for WM-AU. Shifting what seems to be an initial science focus to encompass humanities and social sciences is also an exciting challenge."

–Pru Mitchell,
Manager, Information Services, Australian Council for Educational Research
President, Wikimedia Australia

❧

"WikiCite is an incredibly important project not just for academics and scholars but for the entire world. Creating universal, shareable citations has implications for the chain of verification needed for a modern information and knowledge ecosystem. Only a group of people with the ethos and motivation of the Wikimedia community could fulfill the mission of creating a citation system that spans geographic and pedagogical boundaries. Where financial incentives fail, the free and open community rises to the challenge."

–Andrew Lih
journalist, author of The Wikipedia Revolution

❧

"Discovering relevant sources on a topic is a major part of a researcher's workflow. As librarians, we teach students and researchers to follow references to explore the literature and to take care of bibliographic details during referencing. These tasks rely on the availability of citation data and only open citation data enables transparent and

reproducible bibliometric analyses to understand the impact of scholarship. I participated in the WikiCite 2017 meeting on behalf of my library, motivated by these topics, and was astonished to see what this community has achieved since the meeting in 2016."

–Philipp Zumstein
Mannheim University Library, LOC-DB project, Zotero contributor

Financial report

The Wikimedia Board of Trustees approved the WikiCite 2017 initiative as a recipient of restricted grants from 3 funders: the Alfred P. Sloan Foundation, the Gordon and Betty Moore Foundation, and the Science Sandbox initiative at the Simons Foundation. Dario Taraborelli and Jonathan Dugan, acted as co-PIs on the grant proposal and managed the WikiCite budget, in coordination with the organizing committee. As of October 31, 2017, the funds have been used as follows:

Total funding	$65,869	Grants , including a $5,869 balance carried forward from WikiCite 2016
Total spent	*($66,949)*	See cost breakdown below
Balance	*($1,080)*	

Cost breakdown

Travel scholarships	*$34,437*	We issued travel scholarships to allow 39 participants with no additional sources of funding to attend the event
Venue	*$22,609*	Conference rooms and facilities; catering; accommodation for participating Wikimedia Foundation staff
Dinners	*$4,210*	Conference dinners
Other costs	*$5,694*	Administration costs; other outreach travel costs

(Reporting period: November 1, 2016 - October 31, 2017)

Acknowledgments

WikiCite 2017 was organized by: Jonathan Dugan, Rachel Farrand, Daniel Mietchen, Cameron Neylon, Lydia Pintscher, Elizabeth Seiver, Sarah R. Rodlund, Dario Taraborelli.

The initiative was made possible by generous funding from the **Alfred P. Sloan Foundation**, the **Gordon and Betty Moore Foundation**, and the **Science Sandbox initiative at the Simons Foundation**. We're grateful for their support.

The organizing committee also wishes to thank **Wikimedia Austria** and the **Wikimedia Foundation** for their support, in particular: Claudia Garad, Zhou Zhou, Stephen LaPorte, Caitlin Virtue, Jonathan Curiel, Tony Le, Alan Lau, Meron Kristos, Pamela Swaby.

A Wikidata tutorial explains how to create a bibliographic entry • Wikimedia Argentina, CC BY SA 4.0