Research Ideas and Outcomes 5: e35820
doi: 10.3897/rio.5.e35820

Grant Proposal

Robustifying Scholia: paving the way for knowledge discovery and research assessment through Wikidata

Lane Rasberry[‡], Egon L. Willighagen[§], Finn Årup Nielsen[|], Daniel Mietchen[‡]

‡ Data Science Institute, University of Virginia, Charlottesville, United States of America
§ Dept of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, Maastricht, Netherlands
| Technical University of Denmark, Kongens Lyngby, Denmark

Corresponding author: Daniel Mietchen (daniel.mietchen@virginia.edu)

Reviewable v1

Received: 29 Apr 2019 | Published: 02 May 2019

Citation: Rasberry L, Willighagen E, Nielsen F, Mietchen D (2019) Robustifying Scholia: paving the way for knowledge discovery and research assessment through Wikidata. Research Ideas and Outcomes 5: e35820. https://doi.org/10.3897/rio.5.e35820

Abstract

Knowledge workers like researchers, students, journalists, research evaluators or funders need tools to explore what is known, how it was discovered, who made which contributions, and where the scholarly record has gaps. Existing tools and services of this kind are not available as Linked Open Data, but Wikidata is. It has the technology, active contributor base, and content to build a large-scale knowledge graph for scholarship, also known as WikiCite. Scholia visualizes this graph in an exploratory interface with profiles and links to the literature. However, it is just a working prototype. This project aims 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 server throttling and handling of large or incomplete datasets. Further goals include integrating Scholia with data curation and manuscript writing workflows, serving more languages, generating usage stats, and documentation.

Keywords

Wikidata, WikiCite, Scholia, bibliographic metadata, knowledge graphs, bibliometrics, scientometrics, SPARQL, research infrastructure

Preface

This document represents a slightly edited version of the original proposal that was submitted to the Alfred P. Sloan Foundation on February 8, 2019. The original proposal then underwent peer review, as a result of which the proposed project "Robustifying Scholia" was funded. We plan to document this review process as well.

In comparison to that original proposal, the current version differs mainly in that it has an abstract (taken from the cover letter) as well as more space for references, tables, and figures.

What is the main issue, problem, or subject and why is it important?

This project seeks to strengthen the infrastructure behind using Wikidata (Vrandečić 2012) to mobilize Linked Open Data for knowledge discovery in scholarly and other contexts, as well as for research assessment.

Linked Open Data is a key element in putting the idea of a semantic web into practice (Bizer et al. 2009), and Wikidata is the channel through which anyone can access or curate such data in their browser. Scholia provides a set of windows into Wikidata's scholarly content by presenting subject-specific sets of visualizations for common queries useful to researchers (Nielsen et al. 2017). The Scholia team expects the tool's popularity to grow, particularly due to its integration with the consistently popular Wikimedia platforms.

While the marketplace offers a range of services with overlapping functionality, none of the available options are community-curated, most use proprietary code, and those few that are based on free and open-source software make use of non-open data, which impedes further aggregation and reuse. We strive to apply Wikimedia values of openness to this problem by developing Scholia as an entirely free and open alternative to the competition, focusing on collaborative curation rather than seeking to capture and contain user contributions.

The major problem with Scholia at the moment is that it is a beta prototype which works, but has not had the necessary development to make its infrastructure ready for Wikimedia-scale use. In this project, we seek to develop Scholia into version 1.0 for pilot communities in anticipation of the approaching opportunity to adapt it for use on a global scale for all areas of research (Lemus-Rojas and Odell 2018).

Table 1 provides an overview of the kinds of visualizations (Fig. 1) which Scholia can present. It does so by querying Wikidata for information of scholarly interest about the concept to be profiled (Malyshev et al. 2018) and matching that with related information from Wikidata, as well as content from ist sister sites, Wikimedia Commons and Wikipedia. The figure highlights cases which work, but a range of challenges can cause these visualizations to break. By addressing the underlying basic problems, we hope to bring enough stability to Scholia that the Wikimedia infrastructure and community can more fully begin to engage with and participate in discussions around scholarly data.

Table 1.

Sample Scholia Visualizations. Scholia creates scholarly profiles by presenting the output of standardized sets of queries over the Wikidata corpus. Popular query sets include researchers, topics, and institutions. In some specific cases, e.g. to visualize large coauthor networks or the entire academic output of large universities, the current query results do not compute or render properly due to technical limitations which this project is to address. In the submitted version of the proposal, the table included a miniature version of the images in Fig. 1 as well as of the complete *author* profile referred to in the legend to Fig. 1a.

To visualize this concept in a Wikimedia profile...	**Scholia presents a data visualization like this...**
Topics about which a journal, researcher, or institution publish most often	Fig. 1a
Counts of publications from a person or organization by year	Fig. 1b
Networks of co-occurring topics in research, or of clusters of co-authors	Fig. 1c
Locations of research, or of institutions active in a field of research, or groups which receive a type of funding	Fig. 1d
Timelines of a researcher's institutional affiliations, or the history of research around a topic	Fig. 1e
Charts ordering all sorts of popularity counts, like most cited papers, researchers, or institutions for a topic	Fig. 1f

This project seeks to assemble key data corpora for narrow use cases in pilot communities to get deep feedback on the performance, design, accessibility and other features of Scholia's most important functionalities. The near future of Scholia's development will require some technical decisions in database architecture, hardware investment and planning for institutional partnerships. Before making those foundational decisions and doing Wikimedia-scale outreach, the team seeks to pilot, assure partner buy-in, and establish community values for the project.

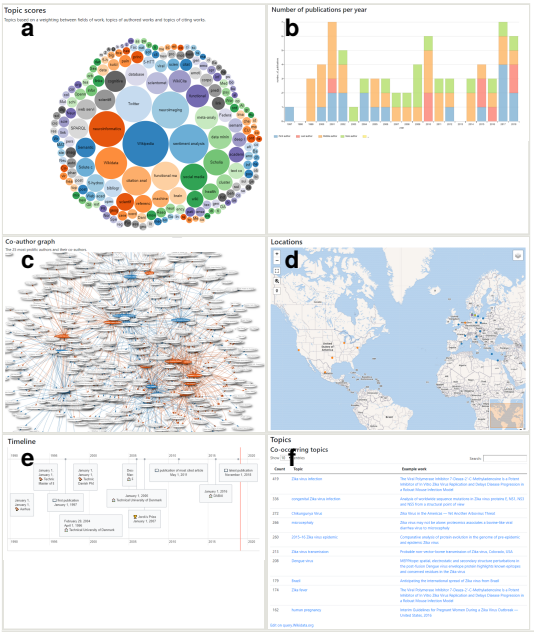

Figure 1.

Screenshots of Scholia with examples of the kinds of visualizations it provides. Such individual visualization panels are then combined in a predefined way into profiles for authors, topics, organizations, works, events, locations or other units of interest. The examples are taken from the Scholia *author* profile for "Finn Årup Nielsen", available via https://tools.wmflabs.org/scholia/author/Q20980928, and the Scholia *topic* profile for "Zika virus", available via https://tools.wmflabs.org/scholia/topic/Q202864. All of these images were made available through Wikimedia Commons as part of the drafting process for this proposal. The files' locations there are indicated, and they are all licensed CC0/Public Domain.

a: "Topic scores" panel for an *author* profile. The image is available as Scholia_-_January_2019_-_Finn_%C3%85rup_Nielsen_as_author_-_topic_scores.png , and a snapshot of the complete author profile as Scholia - January 2019 - Finn Årup Nielsen as author -.png. doi

b: "Number of publications per year" panel for an *author* profile. The image is available as Scholia_-_January_2019_-_Finn_%C3%85rup_Nielsen_as_author_-_publications_per_year.png. doi

c: "Co-author graph" panel for a *topic* profile. The image is available as Scholia_-_Zika_virus_as_topic_-_January_2019_-_co-author_graph.png , and a snapshot of the complete topic profile as Scholia - Zika virus as topic - January 2019.png. doi

d: "Locations" panel for an *author* profile, showing geolocations connected to the author's information in Wikidata. The image is available as Scholia_-_January_2019_-_Finn_%C3%85rup_Nielsen_as_author_-_locations.png. doi

e: "Timeline" panel for an *author* profile. The image is available as Scholia_-_January_2019_-_Finn_%C3%85rup_Nielsen_as_author_-_timeline.png. doi

f: "Co-occurring topics" panel for a *topic* profile. The image is available as Scholia_-_Zika_virus_as_topic_-_January_2019_-_co-occurring_topics.png. doi

What is the major related work in this field?

Competing products either restrict data export or require software installation to present data visualizations of the kind that Scholia provides in any modern browser using just JavaScript and Wikidata.

Products which perform comparable functions to Scholia include Elsevier's Scopus and SciVal (which visualizes Scopus data), Clarivate Analytics' Web of Science and InCites (which visualizes Web of Science data), ResearchGate, Google Scholar, Digital Science's Dimensions and Microsoft Academic Search. Each of these products has its own ecology of features, investments valued to tens of millions of dollars, and restrictions on data reuse (partly because they aggregate from other sources that have restrictive terms). There are also open-source software packages like VIVO (Börner et al. 2012), VOSviewer van Eck and Waltman 2009) or SciMAT (Cobo et al. 2012) that provide scholarly profiles and visualizations of scientific networks, or Open Knowledge Maps that cluster publications on a given topic but their data are often sourced from the commercial products and thus not licensed for unrestricted reuse.

In comparison, Scholia commits entirely to being free and open regarding both its software (which is available under the GNU General Public License, version 3) and data (which is all based on Wikidata and hence CC0/Public Domain). This commitment to openness makes Scholia the only product which (1) curates only public domain data that anyone can reuse for any purpose, including exporting into other products, e.g. as envisaged in the VIVO-based ROSI project (Hauschke et al. 2018), (2) channels collaborative curation efforts back onto the data it is based on, (3) uses only free and open-source software, (4) integrates into Wikimedia projects and anything which accesses Wikimedia content pipelines, (5) fully aligns with the Open Movement and sets standards for openness.

Wikipedia has defined minimum standards for access to information. Similarly, Scholia will not compete with the most costly features of alternative products, but instead seeks to raise the world's minimal expectations regarding knowledge discovery, research assessment, and related activities.

Why is the proposer(s) qualified to address the issue or subject for which funds are being sought?

The proposers are qualified for including the three people who have established Scholia as a working beta tool, for having the academic background to credibly share Scholia with research institutions, and for already having insider credibility as community members within the larger WikiCite project Mietchen et al. 2017) and Wikimedia platforms more broadly.

Complementary to Scholia itself, the ecosystem of media around the tool includes the team's own documentation of Scholia's code, usage instructions, conference presentations, and presentation in academic literature. The published record also includes media from Wikimedia community discussions about Scholia, the testimony of academic partners who have piloted Scholia use of their own accord, and the market research descriptions which list Scholia among key Wikimedia development trends.

The core team consists of Daniel Mietchen, a biophysicist and data scientist at the Data Science Institute at the University of Virginia, along with Finn Årup Nielsen, a computer scientist and neuroscientist in DTU Compute at the Technical University of Denmark who started Scholia, and Egon Willighagen, a chemist in the Department of Bioinformatics at Maastricht University.

What is the approach being taken?

The Scholia project uses the Wikimedia custom of agile development, in which each improvement is an incremental change which makes a permanent impact both in the function and published record of the product. The changes this project proposes are independent of each other, and will be addressed in parallel and deployed into the working tool without disrupting its user base.

The goal of such infrastructure development is to create channels for a community-crowdsourced feedback loop of contribution of content, comments, criticism and improvements. As with any Wikimedia project, the large majority of the content development and labor which improves Scholia comes from users with no direct relationship to the core team of developers. Each of the individual plans for development seeks to lower a barrier which users must overcome to participate in the workflow of using Scholia and contributing content to develop it further.

1. Back-end development and testing with pilot corpora

This project requires systematic testing of Scholia performance across possible use cases or usage scenarios. On that basis, we will assemble a corpus of examples that test Scholia's technical limits that can help us optimize the infrastructure or inform technical design decisions. We will make such decisions around the types of visualizations available through Scholia and how they are cached or preserved, the ways in which the data to visualize gets into Scholia from Wikidata, the ways in which users can configure the experience (e.g. for comparisons), and the ways in which Scholia is integrated with WikiCite curation workflows, or hardware requirements.

While these technical test sets may be of limited use or interest outside the Scholia development team, the systematic testing of Scholia's limits can also help identify circumstances where the tool works well, and in conjunction with usage information, we can then start to build pilot datasets like the Zika corpus or the Invasion biology corpus to serve as examples that engage different user communities. Having example sets to show

off creates a model and workflow for others to emulate to open, expand, integrate or clean up the datasets which are relevant to them.

2. Redesign of the Scholia user interface for better usability, e.g. site navigation and internationalization

The existing beta version of Scholia is functional but requires integration with Wikimedia standards of high usability at a global scale and using accessible design whenever possible. Wikimedia projects already have multilingual infrastructure into which we can integrate Scholia to share its language interface with other Wikimedia translation efforts, particularly through Wikidata.

This project will seek review from a user experience professional to ensure that Scholia meets contemporary standards for usability, including accessibility in design.

3. Improving integration with WikiCite curation workflows, e.g. around missing data

When information is lacking in Wikimedia projects, signals such as "citation needed" invite users to contribute, enrich, and critique the available information. In the Wikimedia way, as Scholia presents and visualizes information, it also identifies content gaps.

Currently, Scholia is primarily a tool for visualization of Wikidata content. Separately, the WikiCite project has tools for data curation on Wikidata, such as the Author Disambiguation tool (Smith 2019) to replace the strings of unidentified author names with the Wikidata identifiers of the respective authors, or the SourceMD tool (Manske 2019) for ingesting source metadata. We propose to integrate such popular existing tools more closely with Scholia, so that as with Wikipedia, anyone who is reading the information gets an invitation to add or modify content.

4. Enhancing Wikidata-based reference management for scholarly writing workflows

Besides the visualizations, Scholia has a number of additional functionalities e.g. regarding entity recognition or reference management. The latter is due to a Python library that processes BibTeX and thereby provides the ability to cite references in TeX/LaTeX documents through their Wikidata identifier or DOI, such that the citation is generated based on the reference's metadata as available from Wikidata.

This mechanism, while functional, needs to be made more comprehensive and robust. If that is achieved and its usage scaled up, this would provide a compelling way for the community of BibTeX users to share the curation of their metadata through Wikidata.

We currently have no plans to expand this functionality beyond TeX/LaTeX for other writing environments, but a related JavaScript library, Citation.js (Willighagen 2019), is being developed with input from the Scholia team, and so is pandoc-wikicite, a filter for the open-source document converter tool Pandoc (Voß 2019). Both libraries can handle formats beyond BibTeX and Wikidata and could help expand a collaborative way of reference metadata management beyond the TeX community. Zotero has also been integrated with Wikidata in a way that allows data exchange in both directions. It can ingest and output

files in a variety of formats including BibTeX, so it can serve as a bridge between different writing environments. Another important development in this space is that Citoid, a MediaWiki extension that facilitates reference management on Wikimedia sites other than Wikidata, is scheduled to be adapted to Wikidata in the coming months, which will provide further integration with Wikimedia writing workflows.

5. Establishing metrics to generate usage stats for Scholia pages and key bibliographic properties and items

We already publish some basic metrics for usage statistics of Scholia-related Wikidata properties (cf. Table 2). We lack more granular insights, like the impact of Scholia on a particular field or institution. Paving the way towards routine availability of such metrics is thus an aim of this project, so that we support our contributors and partners with the media metrics that institutions use to demonstrate the value they get from engagement. What is clear is that the traffic to Scholia pages is increasing (cf. Fig. 2).

Table 2.

Snapshot of the live statistics panel on the Scholia homepage as of 7 February 2019, with some key stats regarding Wikidata content used by Scholia. A corresponding screenshot is available on Wikimedia Commons, and so is a timeline for some of these stats.

Count	Description
7063397403	Total number of triples
174306082	Citations
91891486	Author name strings on items about works
17773585	Items with a PubMed ID
16440023	Items with a DOI
7236781	Items with a geolocation
5876764	Links from items about works to items about their authors
5382376	Links from items about works to items about their main subjects
4613688	Links from items about works to items about their main subjects
2559260	Items with a taxon name
452547	Items about authors with an ORCID profile that has public content

6. Improving documentation of corpus, code, queries, workflows, examples and related resources, as well as limits of Scholia

Documentation of Scholia's content, code, and user instructions is significant beyond Scholia and models and teaches the concept of openness in general. By prioritizing documentation, we also establish a historical record of values. Universities, libraries, and the public should expect and protect the level of transparency, accessibility, and openness

for which we are setting a standard. Our documentation priorities include routine usage instructions, lay and accessible interpretations of visualizations, notes on the underlying queries and data sources, and statements about gaps and biases.

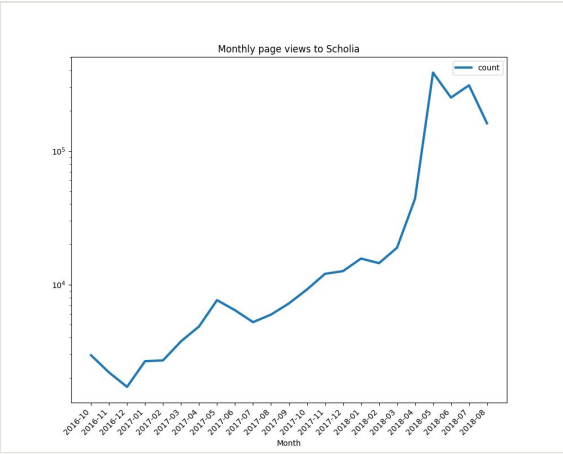

Figure 2. doi

Visualization of Scholia traffic logs (note the semi-logarithmic scale). This data has not been analyzed yet, so cannot be used to draw any conclusions other than that usage grows. Multiple contributing factors seem likely, including web crawlers, generic growth of Wikidata and WikiCite content, increased interlinking both within Wikidata and between Wikidata and other websites, especially Wikimedia projects, as well as WikiCite or Wikidata outreach activities, which often feature Scholia prominently. From Nielsen and Mietchen 2019, CC BY 4.0.

What will be the output from the project?

The main output of this project are technical improvements to Scholia–which will result in the release of version 1.0–as a Wikimedia-based tool for presenting scholarly profiles based on Linked Open Data from Wikidata. Users can view these profiles to gain insights of general or specialist interest to answer questions and guide future research.

If we are successful to the limits of our dreams, then the Wikidata community will play a significant role in shaping the Data Science Revolution in the context of scholarly research, knowledge discovery and research evaluation, and Scholia will be used routinely by students, researchers, journalists and many others around the globe and across disciplines. If we fall short, then we will learn about technical, social and other factors that need to be addressed before such dreams can become reality, and identify some niches where Scholia might find more favourable conditions.

Scholia is a tool which people are using already. Building on the growing awareness of Wikidata in cultural heritage contexts, including as a Linked Open Data hub for libraries,

pilot projects have begun to explore systematically the roles Wikidata could play in library-related workflows. Some of these pilots focused on integrating Wikidata into production library systems, others on offering library services around scholarly profiles for faculty.

For the personal profile use case (which was at the origin of Scholia), the tool has been demoed at institutions in the US, the UK, the Netherlands, Germany and Spain and was highlighted in a white paper of the Association of Research Libraries' Wikidata Task Force (Allison-Cassin et al. 2019). Some Wikimedia-external websites are linking to Scholia profiles, e.g. the Journal of Cheminformatics, and the Scholia tool suite has been used to both write and track publications about Scholia, while researchers outside the Wikimedia community have begun to cite Scholia in their work about academic profiling, author disambiguation and related subjects.

Links to Scholia pages have also been integrated into a number of other Wikimedia tools, e.g. Reasonator, Wikidata Hub, the Author Disambiguator (to which Scholia links back, so as to smooth data curation workflows) as well as the MediaWiki templates Wikidata infobox and Scholia. Through such templates, Wikimedia Commons and several Wikipedias (e.g. Basque, Swedish, English) now have mechanisms to link to relevant Scholia profiles, and we have begun to track some of this usage in a rudimentary fashion. The large user bases of these wikis provide a potential for future growth of Scholia usage, and while eliciting such growth will not be in the focus of the current project, the technical work undertaken in its framework will have reuse at Wikimedia's massive global scale in mind.

What is the justification for the amount of money requested?

Scholia has been in beta testing since late 2016. In that time, various pilot communities have flagged technical problems and made feature requests in the tool's GitHub repository and elsewhere. The Scholia team seeks to resolve issues to end beta testing, release Scholia 1.0, and plan for the next stage of the tool's integration into the Wikimedia platforms.

The team seeks to be cautious in this project, and rather than funding any major new visible features, make its back-end infrastructure stable, well designed, well documented, and orderly for others to test and examine. On the front-end, the priority is conforming to Wikimedia standards for accessibility and internationalization.

All of the staff allocations (cf. Table 3) will have some input into all aspects of this phase of development, but roughly, most of the development will go to back-end infrastructure, less will be for front-end design and accessibility, and less still will go to engaging with end users including meeting UX standards, recruiting testing from pilot communities, and documentation for the project.

Table 3.

Categories of Robustifying Scholia expenses. For the detailed budget, see Suppl. material 1.

expense	~% **of budget**	**description**
3 investigators, 0.1 time each	11%	oversight and administration
back-end developer	25%	add features and improve function
front-end developer	13%	apply interactive wiki interface
UX designer	5%	accessibility
community outreach	5%	user feedback throughout development
documentation / student research	6%	test workflows and publish instructions
other direct costs	5%	publishing and travel
benefits	17%	defined by university
overhead	13%	defined by university
total	100%	

What other sources of support does the proposer have in hand or has he/she applied for to support the project?

Scholia has never had dedicated financial support for development. The team has no plans to seek additional funding for any of the features for which this project seeks funding. Scholia's origins are as an unfunded side project of the core team, in the general reuse of open software and open data, and in crowdsourced Wikimedia community engagement.

The most valuable existing support which the project has is social and technical integration with the Wikimedia platforms and the Wikimedia community. However, part of the writing for this proposal has been accomplished by Daniel Mietchen and Lane Rasberry on staff time at the Data Science Institute of the University of Virginia.

Data in this project will have primary publication in Wikidata and is dependent on the infrastructure of that project. Scholia code is hosted on GitHub, with archival copies being released on Zenodo on a regular basis (see example). The code is run through the Wikimedia Toolforge. Many of Scholia's features are features in the Wikidata platforms, hence as Wikidata develops, so do these features become better in Scholia. Many features in the Wikimedia ecosystem are interrelated to others, and this project operates in that continually developing environment.

What is the status and output of current and/or previous Sloan grants?

This project has not received previous Sloan funding. The organizers of this project have no previous Sloan grants. In the Wikimedia ecosystem, Sloan has funded other projects and the Wikimedia Foundation itself with indirect connections to this Scholia project. The difference is that the Wikimedia Foundation is a platform which provides a space for community organization, and Scholia is one of the community projects which operates within both the Wikimedia platforms and the general public commons.

In 2017, Sloan provided funding to the Wikimedia Foundation for a project to convert metadata for media files, mostly images on Wikimedia Commons, into open linked data. That project is in a separate domain than Scholia, and the Wikimedia Foundation is its own unrelated entity.

Annually from 2016-18, Sloan has sponsored the Wikimedia Foundation to organize the WikiCite conference, which is an event for about 100 people to discuss the curation of source metadata in Wikidata. Our Scholia project depends on WikiCite data and seeks to contribute more data to the collection discussed at that conference. However, the Wikimedia Foundation and its conference are separate and independent from this proposal and Scholia's tool and content development, although this proposal's principal investigator Daniel Mietchen has been an organizer for the WikiCite conference.

Appendix

Conflicts of Interest / Sources of Bias

Principal investigator Daniel Mietchen and project advisor Dario Taraborelli have a conflict of interest for being co-organizers of the Sloan-sponsored WikiCite conference. The other core organizers and the grantee institution have no conflict of interest.

Scholia is a free and open community project. Anyone may participate as an end user or by registering a Wikimedia account and joining discussions. In Wikimedia projects, there is no identified precedent of a conflict of interest problem in a project comparable to Scholia. The Scholia team will watch for conflict of interest as is customary in Wikimedia projects, and will report any issues which arise, but does not anticipate undue participation of any conflicted stakeholders despite the openness of the project.

The Scholia team expects to have the participation of individuals who make purchase decisions for knowledge discovery tools for their institutions. The Scholia team does not anticipate that seeking input and participation from people at this level will raise the challenges of a conflict of interest. Such people could include scholarly communication librarians and human resources teams at research institutes.

Because Wikidata ingests external data, it also ingests the bias of the sources of that data and its environment. Visualizations like those provided by Scholia can help identify such biases. Fig. 3 highlights counts of 2018 geolocation data showing that Wikidata's content better covers some locations over others. Various commentators and researchers discuss systemic bias in Wikimedia platforms. This project aims to counter Wikimedia biases by seeking out underrepresented data sets in the pilot corpora.

Figure 3. doi

Map of geographical bias in Wikidata (by Adam Shoreland, via Wikimedia Commons, CC0). The color indicates how many Wikidata items have a geolocation within the corresponding pixel, with brighter colors indicating a larger number. The observed biases are a combination of multiple factors, including human settlement patterns and Internet access. In a similar fashion, Wikidata can be used to visualize other kinds of biases, across data types, domains of knowledge, or historic eras.

Attention to Diversity

This project starts with two mandates for diversity: that of the Wikimedia community and that of the University of Virginia as project host. These mandates require diversity in the appointment of funded positions, and in recruitment of community engagement in the development of the project, and in targeting the pilot communities who will be among the first beneficiaries of project outcomes.

In funded roles, the project starts with 3 investigators in 3 different countries who will develop the tool in their 3 different native languages and additionally English. Later appointments will seek other dimensions of diversity. The project already has collaborations with established Wikimedia community organizations to provide user feedback during development, including groups organized by country, organizations for gender and racial diversity, or inclusivity of specific academic fields.

Scholia itself is a tool which can identify academic accomplishments of minority groups and their members. For example, scientists who are ethnic minorities, women, or LGBT+ often seek to publicly identify themselves to be counted and encourage more people of their

demographic to join the sciences. Some queries which Scholia's native infrastructure can accomplish, but which are too progressive and provocative for competing products, include "ratios of scientists by ethnicity at a given university" (prototype) or "academic fields sorted by the count of LGBT+ researchers publishing on the topic" (prototype).

Information Products Appendix

This project seeks to be a model of Wikimedia openness in all information product outputs. Every information product which this project creates will be aligned with the Wikimedia ideal of free media and have compatibility with the appropriate Wikimedia project licenses, which are CC0 for data, CC BY or CC BY-SA for most media and text, and free and open software licenses to operate on Wikimedia servers.

This project will present datasets, software, documentation, and the published text of online community discussion as part of the primary goal of developing Scholia as an online tool for exploring the Wikidata knowledge graph of WikiCite data. We will put data produced in this project into the Wikidata platform which offers various format options for anyone to export their own copy of the content. Beyond applying open licenses to the primary information products, this project additionally seeks to be open in development, community participation, and public discussion around the project. These processes and conversations will also happen in the open in ways that create media records with open licenses which anyone can access or scrutinize.

To increase accessibility to information products beyond the Wikimedia platforms, we will mirror the publication of some products in more traditional spaces. Examples of additional distribution plans include using GitHub as a code repository for this project and Zenodo for archival copies to make these resources more accessible.

This project will reuse code and content whenever possible, always with a Wikimedia compatible open license. The policy which best describes constraints on this project are the Wikimedia policies on openness, such as their Open Access Policy.

Everything produced by this project will be accessible online for anyone to access without paying a cost to access, export, remix, or reuse it.

Acknowledgements

The authors would like to acknowledge the WikiCite community's contributions to bibliographic and related data in Wikidata and to Scholia's documentation and code.

Funding program

The project is funded by the Alfred P. Sloan Foundation under grant number G-2019-11458.

Grant title

Robustifying Scholia: paving the way for knowledge discovery and research assessment through Wikidata

Hosting institution

Data Science Institute, University of Virginia

Ethics and security

Information on ethics or security was not required but we plan to explore these issues nonetheless.

Author contributions

All four authors were involved in conceptualizing the project. LR and DM wrote it up.

References

- Allison-Cassin S, Armstrong A, Ayers P, Cramer T, Custer M, Lemus-Rojas M, McCallum S, Proffitt M, Puente M, Ruttenberg J, Stinson A (2019) ARL White Paper on Wikidata: Opportunities and Recommendations. https://www.arl.org/storage/documents/publications/2019.04.18-ARL-white-paper-on-Wikidata.pdf. Accessed on: 2019-4-27.
- Bizer C, Heath T, Berners-Lee T (2009) Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 5 (3): 1-22. https://doi.org/10.4018/jswis.2009081901
- Börner K, Conlon M, Corson-Rikert J, Ding Y (2012) VIVO: A Semantic Approach to Scholarly Networking and Discovery. Synthesis Lectures on the Semantic Web: Theory and Technology 2 (1): 1-178. https://doi.org/10.2200/s00428ed1v01y201207wbe002
- Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F (2012) SciMAT: A new science mapping analysis software tool. Journal of the American Society for Information Science and Technology 63 (8): 1609-1630. https://doi.org/10.1002/asi.22688
- Hauschke C, Cartellieri S, Heller L (2018) Reference implementation for open scientometric indicators (ROSI). Research Ideas and Outcomes 4 https://doi.org/10.3897/rio.4.e31656
- Lemus-Rojas M, Odell J (2018) Creating Structured Linked Data to Generate Scholarly Profiles: A Pilot Project using Wikidata and Scholia. Journal of Librarianship and Scholarly Communication 6 (1). https://doi.org/10.7710/2162-3309.2272
- Malyshev S, Krötzsch M, González L, Gonsior J, Bielefeldt A (2018) Getting the Most out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph. In:

Vrandečić D, Bontcheva K, Suárez-Figueroa MC, Presutti V, Celino I, Sabou M, Kaffee L, Simperl E (Eds) LNCS., 11137. 17th International Semantic Web Conference (ISWC'18), Montery, USA, 8–12 October 2018. Springer

- Manske M (2019) SourceMD. BitBucket. Release date: 2019-4-15. URL: https://bitbucket.org/magnusmanske/sourcemd
- Mietchen D, Taraborelli D, Nielsen FÅ, Waagmeester A (2017) WikiCite: Citations needed for the sum of all human knowledge. Figshare https://doi.org/10.6084/M9.FIGSHARE.5306122.V1
- Nielsen FÅ, Mietchen D, Willighagen E (2017) Scholia, Scientometrics and Wikidata. Lecture Notes in Computer Science 237-259. https://doi.org/10.1007/978-3-319-70407-4_36
- Nielsen FÅ, Mietchen D (2019) Scholia as of November 2018. Zenodo https://doi.org/10.5281/zenodo.2653163
- Smith A (2019) author-disambiguator. GitHub. Release date: 2019-3-18. URL: https://github.com/arthurpsmith/author-disambiguator
- van Eck NJ, Waltman L (2009) Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84 (2): 523-538. https://doi.org/10.1007/s11192-009-0146-3
- Voß J (2019) wcite. GitHub. Release date: 2019-4-21. URL: https://github.com/wikicite/wcite
- Vrandečić D (2012) Wikidata. Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion https://doi.org/10.1145/2187980.2188242
- Willighagen LG (2019) Citation.js: a format-independent, modular bibliography tool for the browser and command line. PeerJ Preprints https://doi.org/10.7287/peerj.preprints.27466v1

Supplementary material

Suppl. material 1: Budget for "Robustifying Scholia" doi

Authors: Lane Rasberry, Egon L. Willighagen, Finn Årup Nielsen, Daniel Mietchen
Data type: budget spreadsheet
Brief description: The file contains the budget that was submitted along with the proposal.
Download file (18.15 kb)