Scholia for Software

Lane Rasberry[‡], Daniel Mietchen[§,|]

‡ University of Virginia, Charlottesville, United States of America
§ Ronin Institute of Independent Scholarship, Montclair, United States of America
| Institute for Globally Distributed Open Research and Education (IGDORE), Jena, Germany

Corresponding author: Lane Rasberry (lr2ua@virginia.edu), Daniel Mietchen (daniel.mietchen@ronininstitute.org)

Abstract

Scholia for Software is a project to add software profiling features to Scholia, which is a scholarly profiling service from the Wikimedia ecosystem and integrated with Wikipedia and Wikidata. This document is an adaptation of the funded grant proposal. We are sharing it for several reasons, including research transparency, our wish to encourage the sharing of research proposals for reuse and remixing in general, to assist others specifically in making proposals that would complement our activities, and because sharing this proposal helps us to tell the story of the project to community stakeholders.

A "scholarly profiling service" is a tool which assists the user in accessing data on some aspect of scholarship, usually in relation to research. Typical features of such services include returning the biography of academic publications for any given researcher, or providing a list of publications by topic. Scholia already exists as a Wikimedia platform tool built upon Wikidata and capable of serving these functions. This project will additionally add software-related data to Wikidata, develop Scholia's own code, and address some ethical issues in diversity and representation around these activities. The end result will be that Scholia will have the ability to report what software a given researcher has described using in their publications, what software is most used among authors publishing on a given topic or in a given journal, what papers describe projects which use some given software, and what software is most often co-used in projects which use a given software.

Keywords

research software, open-source software, open science ecosystem, use, re-use, co-use, software citation, software dependencies, Wikidata, altmetrics

Proposal

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

This project seeks to profile research software to help users form a better understanding of the various relationships between scientific software and the research it supports. The main problem is the lack of options for typical researchers to gain insights into the research software ecosystem as a whole. This inability to observe the public commons of software results in lack of understanding of the relationships between software and research. This lack of information causes problems including general inaccessibility of information on critical research tools, lack of credit for software and infrastructure developers, overdependence on under-supported software, lack of appropriate recognition for the tool developers and researchers from underrepresented demographics, and barriers to reproducibility. Basic data that would reduce these problems includes a free and open catalog of software; metadata to identify software provenance, development, and dependencies; demographic data for researcher profiles; and linked data connecting research papers describing software use to identifiers for that software. The ability to query and visualize the ecosystem of research software would surface big-picture context, including ranking tools by usage, identifying under-supported but popular resources, repositioning software contributions to be as worthy of credit as paper authorship, transparency of demographic diversity in creator communities, and making research more reproducible.

What is the major related work in this field?

Wikidata is the major related work in this field on which our project depends, as Wikidata is a general platform for curating and visualizing data in a way that partners with communities and for delivering project outputs to an active user base. For precedents in software identification and classification, various research projects—including the Softcite dataset which matches academic publications to software resources mentioned therein, and the Impact and Diffusion of Open Source Software project which attempts to assign value to open source software—have produced and classified subsets of software catalogs that we will consider as options for developing. Table 1 gives an overview of some of the topics related to the research proposed here:

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

This project has three aspects: Wikidata engagement; software data curation; and diversity recognition. Our team members have experience in each of these areas, and the project overall extends their works in progress for collecting, structuring, and sharing research metadata.

The project team which we proposed includes people with the following skill sets and resources:

1. Wikidata editors who are also Wikimedia community members and familiar with the ethics and norms of the platform
2. Data scientists who are currently engaged in managing metadata of open source software
3. Metadata librarians with general experience in cataloging, including knowledge of contemporary best practices in recording demographic information for individuals
4. Wikidata consultants who provide training to universities and libraries for using Wikidata
5. Wikidata content experts who can execute data uploads when socially supported with ethical review and community approval
6. Software developers who can design and implement new features for Scholia
7. Early and long-time Scholia contributors as project advisors with experience at the intersection of most of what this project proposes
8. Technical writers to produce project documentation

What is the work plan or approach being taken?

Through the data we already have in Wikidata from the WikiCite project, we can write queries to identify, for instance, GitHub users who have co-authored scholarly publications (cf. Fig. 1). This is our starting point from which we will seek more data, develop the software, and organize community conversation.

As this is a Wikimedia platform project, our approach follows Wikimedia community values, including promotion of free and open content, monitoring user engagement to maximize communication impact, recruiting diverse contributors to the project, and opening the project for public discussion. While following these general principles, our approach divides this project into the following workflows:

Collect data about publications linked to software and tools, emphasizing free and open-source software. In order to reduce the size of some of the content gaps around research software we will be mining scholarly repositories—literature repositories like PubMed Central for mentions of software and software repositories like GitHub for mentions of scholarly publications or for CITATION.cff files that support software and data citation.

Register research software and tools through dedicated Wikidata items. The mined data about software will be cleaned, integrated, and converted to Wikidata's data model, subjected to quality checks and then imported into Wikidata. This way, each software or tool gets its own Wikidata item (and thus a dedicated identifier) that will be cross-

referenced with identifiers from other registries. When curating items about research software, we will annotate them with information about versioning, dependencies, file formats, licensing, usage, and software sustainability.

Interlink Wikidata items for software and toolswith other Wikidata items, especially for people and publications. The curation workflows for Wikidata items about research software will be connected with workflows addressing other parts of the Wikidata knowledge graph in ways that allow for further automation. Content development around research software will take place on the basis of pilot corpora, associated Scholia profiles and their curation pages, along with Wikidata quality control and community feedback. The pilot corpora will be assembled such that they overlap and interact with research software in a variety of ways and from different angles. Scholia, Jupyter as well as ImageJ and its ecosystem will serve as initial test cases for research-related software, complemented at later stages by Wikibase and by software for which software management plans are publicly available. Other approaches to building pilot corpora include focusing on specific research domains—starting with biodiversity informatics, cheminformatics, citizen science, computational reproducibility, software ethics, and diversity in tech—or on specific uses of software, e.g. for visualizations, databases, pipelines, or for handling file types important in a particular field. Another way to assemble such corpora will be by following groups of people who create or use software (cf. Fig. 1), amongst which we will prioritize aligned communities like rOpenSci, individuals with an ORCID, those from underrepresented demographics, prize winners, those who published in journals with an open-source policy (like the Journal of Open-Source Software), or who were funded by organizations that have such policies (e.g. Wellcome).

Adapt Scholia profiles for research software and software citation. As a basis for software profiles (see demo for Scholia), we will assemble questions that users could ask about research software. These will be compared with sets of questions that would suit software more generally, for which profiles are already available via WikiDP, albeit without research context (example). Some of these questions may not be readily addressable using present-day Wikidata, as either its relevant data or pertinent data models might be incomplete (see the “Register research software” section above on how this will be addressed). Questions for which the underlying data model is reasonably complete and for which some minimum amount of data exists can then be translated into SPARQL queries and incorporated into Scholia profiles for software. The curation pages of these profiles will also be adapted to better facilitate software citation (similar to article citations) by expanding Scholia's support for CITATION.cff and other citation formats and adding support for CiteAs integration, which would also pave the way for better citation of data and materials. Existing profile types—e.g. for works, authors, journals, and other entity types—will be reviewed from the perspective of incorporating software-related information, and adapted accordingly by incorporating such information or linking to it.

Document and streamline workflows and enhance user experiences. We will document our workflows, so as to enable others to contribute, to reuse and adapt them, and to help scale beyond our pilot corpora. On the way, we will note opportunities to streamline curation workflows or enhance user experience as well as for social and ethical issues that

may arise. By showcasing examples of activities and outcomes and publishing project updates, we will recruit community engagement from Wikidata editors, software developers, and data modelers.

What will be the output from the project?

If we are successful to the limits of our imagination, then we will produce the open data, presentation format, and presentation tool by means of which users will gain satisfying and useful insights into the relationships between software, publications, people, and research. The output of this project will be a catalog of research software linked by structured data to the research ecosystem around it, all visualized through the Scholia tool in the Wikidata project. Table 2 shows outputs for Scholia's end users. Other outputs will include the following:

- A pipeline for mining research software-related information.
- Establishing a Wikidata catalog of research software, including a unique identifier for each entry and a set of queryable descriptors.
- Dataset matching research software to academic papers describing use of that software.
- Software profiles enriched with non-software-related information, and vice versa.
- Documentation of workflows, user experiences, and opportunities for engagement.
- Demographic data for communities of software creators and users.
- Social guidance documentation within the WikiProject for addressing the ethical issues that we or the wiki community identify in this project.

What is the justification for the amount of money requested?

We budget this six-month project to be part-time research by university researchers and students, along with consulting from software developers to complete defined tasks. The labor we plan will be enough to pilot critiqueable software profiles in the Wikimedia platform, so that we can advance the discourse on opening this data, deliver it to the public to get comment and usage metrics, establish fairness and equity as a discourse in this field of development, and set norms for the future where access to this data will be commonplace.

We omit sharing the actual project budget here in this public document. Instead, we report that the proposed budget request was US$130,000, and the budget categories are as described below in Table 3, which shows financial allocations; if we instead describe the budget by category of labor, then our spending is approximately equal parts data processing, community engagement, and software development.

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

We have no other funding in hand but we are exploring collaboration options around our university and beyond.

A common characteristic among all collaborating institutions and individuals in this project is that they are ideological supporters of the Open Movement, either through free and open source software, open science, open data, open licensing, or any other activity which shares digital resources. While open projects themselves often do not have financial or personal resources to share, their open nature makes them reusable resources. We could not realize this project without using open resources which already exist and closely relate to this project's objectives. Resources which this project will use includes datasets from ORCID, GitHub, Crossref, PubMed, and especially Wikidata and the many data resources with which it has already been integrated.

Budget & Detailed Budget Justification

The principal investigator is budgeted full time for 2 of the 6 project months This person will administer the project and manage the Wikidata activities including community discussion, data upload, modeling, and curation. A senior researcher is budgeted 0.5 months to oversee graduate student research into precedents for applying demographic labels at scale in bibliographic databases and developing an ethical recommendation that applies to the software creators and researchers we profile in this project.

Other personnel in the university include two graduate research assistants and two undergraduate data technicians. One of the graduate students will conduct research to promote diversity in stakeholder project impact and ethical demographic labeling, another will do the same but with more outreach communication to Wikidata, and the two undergraduate students will assist with data curation in the Wikidata platform. Pay rates for students will be as recommended by the university student labor union.

Other direct costs include publication fees, which could include open access publishing fees or creation of multimedia communication. Our budget for Scholia software development will go to developers familiar with Scholia and the Wikimedia platform who provide service at market rate.

Project subawards go to academic partners who have open datasets which they will share with this project for the purpose of developing Wikidata content. Subaward partners also each address this project's diversity challenges in their own specific ways.

Commitment to Diversity, Equity & Inclusion

The lead university for this project has university-wide diversity programs which include project recommendations and checklists for compliance. This project will conform to the university's recommendations.

For this project, our team leads are majority non-male. All senior personnel are white and there is no representation from Latino, Black, or indigenous people. We will seek representation from these groups in hiring student researchers and other staffing. Because this project includes demographic research, we need perspectives that can only come from people with lived cultural experiences in those demographic communities. Our hiring strategy includes recruiting through our school's Dean of Diversity, Equity, and Inclusion in our school, who supports projects like ours by ensuring good outreach and a diverse candidate pool.

The research outcomes of this project will include a published research paper on profiling the demographics of researchers. Multiple Wikimedia communities are currently discussing this issue, and any precedent that the Wikimedia community sets has the potential to be an issue of broad interest in discussing DEI in general. To review, our Scholia for Software project will catalog software metadata, which includes cataloging software creators and researchers who publish papers that mention using the software. When we have a collection of names, we can use contemporary data remixing to identify demographic characteristics that match subjects of our biographical research profiles. This could mean generating demographic reports based on gender, ethnicity, LGBT+, or any other reported label when querying sets of software or any arbitrary research team list. Whatever the case, this is a social and ethical issue, and we have research collaborators specialized in biographical databases to oversee student documentation of the discourse in this space. If we are successful in piloting this, we will have established a precedent in methods for surfacing demographic representation in research networks related to software. If we decide not to publish this data in this way, we will at least publish a report that expresses the views of diverse community stakeholders explaining how we made the decision and the risks we identified.

Regardless of the extent to which we apply demographic labels at scale to our bibliographic database, our pilots and documentation will set a precedent in profiling underrepresented demographics, including our target communities of women, black, Latino, and indigenous software creators for our case studies and pilot data collections. In the usual Wikimedia channels of media distribution, we will showcase people from underrepresented demographics so that the general public can have better access to examples of leaders and creators to credit and discuss when considering the major contributors to this field.

Appendix

List of Citations

We are not providing a list of citations, but recommend Scholia as a profiling service for finding scholarly publications which relate to topics of this proposal.

- https://scholia.toolforge.org/

Conflicts of Interest / Sources of Bias

We identify no conflicts of interest in this project. No one involved have present or planned commercial ventures with any of this data.

Bias in the Wikimedia platform is a topic of continuous and multifaceted conversation, and all Wikimedia biases influence this project as well. Popularly discussed biases include underrepresentation of women and minorities; underrepresentation of people, culture, and language in lower and middle income countries; discrimination against certain demographics including people of color regardless of their representation status; and unwanted encroachment of corporate interests despite the Wikimedia community's idealism to favor the public's interest.

Our project cannot counter all biases, but we have plans in place to counter some of it. Our research team selection includes members from underrepresented demographics. When our team develops pilots and examples, we choose case studies which highlight the accomplishments of underserved demographics. Finally, we collaborate with the office of the Dean of Diversity, Equity, and Inclusion in our school to confirm that we are putting sufficient effort towards making our research fair and equitable.

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 at the intersection of research software and 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.

Conflicts of interest

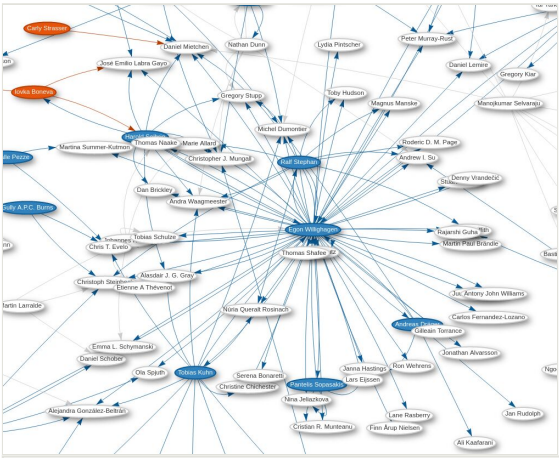

Figure 1.

Scholarly authorship network, filtered for GitHub users, as of 23 September 2021 (source query). The image is available from Wikimedia Commons at http://w.wiki/47JN. The live query results are accessible via https://w.wiki/47JB.

Table 1.

Examples of work related to this project, grouped by general themes.

This project's specific needs	**General background**	**Data curation**
scholarly profiles	scientometrics	data mining
software classification	Linked Open Data	persistent identifiers
linking research to software	research software	entity disambiguation
ethical risks of such data	social machines	social machines
Library cataloging	**Archiving**	**Philosophy**
software cataloging	software sustainability	free and open-source software
version control	computational reproducibility	Wikimedia
software citation	scientific reproducibility	open collaboration
software dependencies	software preservation	
	legacy data	

Table 2.

Input and output of Scholia.

when the end user inputs the name of...	**...then Scholia reports...**	**... and Scholia gives data for...**
a software tool	articles which mention usage	identifiers, classification
a scholarly journal	software mentioned in articles	publication metadata
a researcher	software used	biographical information

Table 3.

Expense categories of the project.

Expense category	~% of budget	description
senior personnel salary	17%	project management
other personnel	16%	data collection and curation
consulting	19%	software development
miscellaneous	7%	community outreach; computation
subaward	21%	content development
personnel benefits	7%	defined by university
administrative costs	13%	defined by university
total	100%	