Research Ideas and Outcomes 7: e66490
doi: 10.3897/rio.7.e66490

Small Grant Proposal

FAIR and open multilingual clinical trials in Wikidata and Wikipedia

Lane Rasberry‡, Daniel Mietchen‡

‡ School of Data Science, University of Virginia, Charlottesville, United States of America

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

Reviewable v 1

Received: 25 Mar 2021 | Published: 25 Mar 2021

Citation: Rasberry L, Mietchen D (2021) FAIR and open multilingual clinical trials in Wikidata and Wikipedia. Research Ideas and Outcomes 7: e66490. https://doi.org/10.3897/rio.7.e66490

Abstract

This project seeks to conduct language translation on metadata labels for research publications, attribution data, and clinical trials information to make data about medical research queriable in underserved languages through Wikidata and the Linked Open Web. This project has the benefit of distributing content through Wikipedia and Wikidata, which already have an annual userbase of a billion users and which already have established actionable standards to practice diversity, inclusion, openness, FAIRness, and transparency about program development. The impact will be localized access to basic research information in various Global South languages to integrate with existing community efforts for establishing the same. Although Wikidata development in this direction seems inevitable, the cultural and social exchange required to establish global multilingual research partnerships could begin now with support rather than later as a second phase effort for including the developing world. Wikipedia and Wikidata are established forums with an existing active userbase for multilingual research collaboration, but the research practices there still are immature. By applying metadata expertise through this project, we will elevate the current amateur development with more stable Linked Open Data compatibility to English language databases. Using the wiki distribution and discussion platform to develop the global conversation about data sharing will set good precedents for the trend of global research collaboration.

Keywords

medical information, clinical trials, Wikipedia, Wikidata, translation, India, East Africa, Hindi, Bangla, Bengali, Swahili

About

The School of Data Science at the University of Virginia received funding through an open call from the Wellcome Trust for this research and development proposal titled "FAIR and open multilingual clinical trials in Wikidata and Wikipedia". This document combines the concept note, full application, and supporting details in that application. Our intent in publishing this is to make the proposal available with free and open copyright licensing for other researchers to reuse and remix as they wish.

The project objective is to import clinical trials metadata from ClinicalTrials.gov into the Wikidata platform, and to curate the data using Wikidata workflows, and also to present the overall process as a model for using the Wikipedia ecosystem to share and remix data. If the project is successful to the extent of our wishes then our hope is that all sorts of researchers and the public will access and use clinical trials data for both traditional and new purposes.

Although we believe that researchers who currently use clinical trials data will find benefit from its curation through the Wikipedia platform development process, we also seek to promote access to this medical information to new audiences both within conventional analysis about clinical trials and in new and unexpected contexts. The new audiences which we anticipate are those whom we already know to browse Wikipedia's medical content, including researchers who prefer to access information outside of English language and non-researchers including students, journalists, and policy makers who previously would not have considered seeking this data were it not accessible through the familiar Wikipedia. By making the data much more accessible and also available for Wikipedia's style of crowdsourcing, we hope that others will develop and reuse this data including by linking trials to papers, people, and organizations; visualizing the trials with charts and maps; general curation of trials by keyword tagging or concept disambiguation; and language translation of technical terms in structured data collections such as Wikidata.

We are sharing this text in alignment with Wikipedia community values of openness and in a contemporary social context where sharing proposals is uncommon but which we wish were more routine. At the time of publishing this proposal we have developed the project but have not yet completed it. This document only presents the proposal, and we will publish our methods, results, and the overall model in a later paper. The term of the project is extended due to COVID-19, and we changed some of the project focus from the original proposal in response to the pandemic. The text here does not account for our response to COVID.

Methodology

1. By default, adopt the established Wikipedia and Wikidata publishing and engagement practices for open, FAIR, documentation, receiving feedback in permanent public forums, and collaboration
2. Contribute to the documentation about the position of Wikipedia and Wikidata in the Linked Open Data ecosystem, particularly emphasizing university participation in import and export of research metadata in the Wikimedia platform and collecting impact metrics for doing so.
3. Within Wikidata, contribute to the WikiCite project which seeks to enrich data around citations and metadata, including PubMed research papers, and subsets of CrossRef, ORCID, and ClinicalTrials.gov. Explore and document possibilities to ingest non-United States clinical trial databases.
4. Identify the set of terms and concepts which are necessary to perform and visualize queries of medical research data, for example, "clinical research sites in a given country with the highest trial completion rates in infectious disease research"
5. Translate those terms to languages including Hindi, Bengali, and Swahili to the level of quality which is established as a norm by existing local community participants in Wikipedia and Wikidata
6. Use Wikipedia and Wikidata's native metrics reporting processes to measure the impact to users and the engagement of peer reviewers

Introduction

The School of Data Science at the University of Virginia received funding through an open call from the Wellcome Trust for this research and development proposal titled "FAIR and open multilingual clinical trials in Wikidata and Wikipedia". This document combines the concept note, full application, and supporting details in that application. Our intent in publishing this is to make the proposal available with free and open copyright licensing for other researchers to reuse and remix as they wish.

The project objective is to import clinical trials metadata from ClinicalTrials.gov into the Wikidata platform, and to curate the data using Wikidata workflows, and also to present the overall process as a model for using the Wikipedia ecosystem to share and remix data. If the project is successful to the extent of our wishes then our hope is that all sorts of researchers and the public will access and use clinical trials data for both traditional and new purposes.

Although we believe that researchers who currently use clinical trials data will find benefit from its curation through the Wikipedia platform development process, we also seek to promote access to this medical information to new audiences both within conventional analysis about clinical trials and in new and unexpected contexts. The new audiences which we anticipate are those whom we already know to browse Wikipedia's medical content, including researchers who prefer to access information outside of English

language and non-researchers including students, journalists, and policy makers who previously would not have considered seeking this data were it not accessible through the familiar Wikipedia. By making the data much more accessible and also available for Wikipedia's style of crowdsourcing, we hope that others will develop and reuse this data including by linking trials to papers, people, and organizations; visualizing the trials with charts and maps; general curation of trials by keyword tagging or concept disambiguation; and language translation of technical terms in structured data collections such as Wikidata.

We are sharing this text in alignment with Wikipedia community values of openness and in a contemporary social context where sharing proposals is uncommon but which we wish were more routine. At the time of publishing this proposal we have developed the project but have not yet completed it. This document only presents the proposal, and we will publish our methods, results, and the overall model in a later paper. The term of the project is extended due to COVID-19, and we changed some of the project focus from the original proposal in response to the pandemic. The text here does not account for our response to COVID.

Who is the project coordinator?

Name: Lane Rasberry
Organization: University of Virginia
Department: School of Data Science
Division: Center for Ethics and Justice
Email: rasberry@virginia.edu
ORCID: 0000-0002-9485-6146

How will you evaluate the success of your activities?

This project will use Wikipedia's own established processes for monitoring and evaluation of university projects to develop and publish general reference information in Wikipedia and Wikidata. We will evaluate this project in these established ways:

1. Content metrics - Report the standard publishing metrics as measured by Wikimedia's own native metrics suite for publishers
2. Diversity and inclusion - Partner with established Wikimedia community organizations; confirm their oversight and approval
3. Impact metrics – Report audience readership as measured by Wikipedia's own native metrics suite for users
4. Quality review - university student researchers will evaluate and publish an evaluation of the source research metadata and the translation process
5. Bias evaluation – we will subjectively publish our opinions on bias we identify and its cause. One obvious source of bias will be availability of open data, as much research indexed in PubMed and ClinicalTrials.gov is not compliant with

recommended metadata standards. This project favors institutions which apply FAIR principles, and we will identify these practices.

Wikipedia as a publishing, technology, and community platform continually introduces processes for content development and evaluation. In 2012, with the establishment of the Wiki Education Foundation, there was a major cultural shift to make Wikipedia compatible with university education and research. Today, that precedent has developed into a suite of open evaluation tools for measuring audience size, levels of engagement, use of fact-checking processes, and a culture applying metrics to perform critical review of Wikipedia's quality. These measurements and processes establish a precedent for this project to follow in doing publishing and content development, operationalizing ethics in digital governance, and publicly demonstrating community conversation in seeking feedback on this project's activities in the context of global Wikimedia content development.

The project coordinator should describe their related research.

My name is Lane Rasberry and I am Wikimedian at the School of Data Science at the University of Virginia. The most important contribution I have made to research has been my advocacy for free and open media. I particularly do this in the context of Wikimedia platform publishing where I showcase Wikidata as a FAIR and open repository which is useful in itself and also a source from which anyone can export information for use in other platforms. For any other project seeking to be FAIR and open, I present Wikidata as a model project to emulate as the standard.

My usual talking point with Wikipedia is that it is the most requested, published, accessed, and consulted source of information in English on nearly every topic, including almost every topic of general interest in medicine. I advocate for institutional partnerships between expert health organizations, typically medical schools and media sources, and the Wikimedia platform. Uniformly all expert health organizations have great challenges in distributing their content to a relevant audience. In comparison, Wikipedia and Wikidata have the world's best available distribution channels and reach a massive audience through Internet search, personal assistants and similar mechanisms. My wish is to match Wikipedia's excellent distribution but lower quality content with the excellent content of expert organizations which have a relatively poorer distribution and media reach. Wikipedia's media reach is useful for any organization with an educational mission to distribute high quality information of general interest. I have managed various programs which integrate information into Wikipedia for the purpose of increased distribution.

Along with that distribution, I have designed policy and published documentation on various ethical considerations of online publishing, including making open media more accessible to diverse audiences who collaborate in friendly online virtual spaces. Such documentation and policy creates an environment where more people can participate and share in the benefits of the FAIR and open global digital commons.

List research projects of the project coordinator.

1. Case studies of the student experience of publishing medical information in Wikipedia
    - Improving the Quality of Consumer Health Information on Wikipedia: Case Series (Weiner et al. 2019)
    - Why Medical Schools Should Embrace Wikipedia (Azzam et al. 2016)
2. Documentation and programming for the Scholia interface to Wikidata and the WikiCite project
    - Robustifying Scholia: paving the way for knowledge discovery and research assessment through Wikidata (Rasberry et al. 2019)
3. Promotion of diversity, inclusion, and friendly collaboration online
    - Automatic Detection of Online Abuse and Analysis of Problematic Users in Wikipedia (Rawat et al. 2019)
    - Queering Wikipedia (Wexelbaum et al. 2015)

How does the project coordinator practice open research?

I Lane Rasberry have been a highly active Wikimedia platform contributor since 2008. I have participated in Wikimedia institutional governance since 2012, including in the platform organizations Wikimedia Medicine, Wikimedia New York City, and Wikimedia LGBT+. I am as active as anyone else in the Wikimedia projects for university programs, the WikiCite project, designing institutional partnerships, and reporting communication metrics for evidence of impact. The Wikimedia platform opens research literature by being the most requested, published, accessed, and consulted source of information on all topics which it covers.

I have been a member of the HIV Vaccine Trials Network Community Advisory Board since 2007. In this role I have advocated for public access to medical research information for a long-running series of clinical trials.

Personally I document and publish notes from organizational meetings of all sorts and publish in the open as a default practice. I advocate for free and open licensing on any media output which anyone intends for distribution to the public.

Who are key collaborators for this project?

School of Data Science, University of Virginia

Daniel Mietchen, data scientist

Lane Rasberry, Wikimedian in Residence

The School of Data Science at the University of Virginia is importing structured data into Wikidata, evaluating the quality of information in Wikipedia and Wikidata, and documenting best practices for university partnerships with the Wikimedia platform. This team identifies the research content for which there is a need in research discovery to be FAIR, available for query, and translated to promote global collaboration.

Daniel Mietchen, data scientist School of Data Science at the University of Virginia, Principle Investigator of the Scholia / WikiCite project to develop the Wikipedia / Wikimedia platform based interface for discovering and visualizing scholarly publications in a free and open system analogous to the popular but closed product Google Scholar. Dr. Mietchen's primary concern are the academic publications, whereas this proposal to Wellcome would integrate clinical trial data into this network and also localize the interface for non-English pilot languages relevant to the developing world. Another way to say this is that Dr. Mietchen operates a Wikipedia-based tool similar to PubMed and Google Scholar, and this project would collaborate with him to integrate ClinicalTrials.gov data into this and permit non-English language use.

UVA Global, University of Virginia

This is a language department at the university where faculty and classes will translate and publish structured data into Wikidata, where it will be FAIR and open in the semantic web.

Wikimedia community organizationsThese organizations provide community feedback on publication in Wikipedia and Wikidata and also on the translation process. These partnerships ensure participation among stakeholders and regional communities of users.

What outcomes will the completed project have?

1. Primary deliverables (Linked Open Data for production)
    1. integrate clinical trials cataloging data from sites like ClinicalTrials.gov into Wikidata
    2. translate a limited vocabulary from ClinicalTrials.gov to Hindi, Bengali, and Swahili
2. Secondary deliverables (cultural products to promote diversity and good ethics)
    1. document Wikidata as a research interface for clinical trials (English, Hindi, Bengali, and Swahili)
    2. publish a general-interest essay on ethical considerations of making medical research data newly accessible

Wikidata is currently the central hub of the open Semantic Web. While various sources such as libraries, repositories, and ClinicalTrials.gov provide "open data", without integration into a Semantic Web portal like Wikidata, the information is largely inaccessible even to field experts. By mapping data to Wikidata, the information becomes accessible to the general public as well as professionals and researchers.

This project will integrate ClinicalTrials.gov into Wikidata, making it once and for all generally machine readable and free and open for export to any other platform. Furthermore, we will translate the search interface to three languages, Hindi, Bengali, and Swahili, bringing this data to those languages for the first time and as a precedent in global diversity.

If we are successful to the limits of our expectations, then all clinical trials data forever after will be free and open in the Semantic Web. Furthermore, we will set the precedent in this project that linguistic diversity must be central to open data projects of global interest. Finally, we will publish a case study of this project in advocacy of accessible open data.

Describe the vision for your proposal, describe how it will promote open research, and explain how you will evaluate impact.

Vision

Our vision for this project is to increase public understanding and global discourse of medical research by making cataloging data on clinical trials much easier to access, query, and visualize in aggregate in English and 3 pilot underserved languages.

Our aims are to enable the following:

1. through publication in Wikidata, professionals in clinical research will have radically increased access to routine data about clinical trials, including from ClinicalTrials.gov and PubMed
2. beyond conventional clinical research data, and for the benefit of the general public and humanities research, through Wikidata we will pilot access to previously inaccessible social ClinicalTrials.gov data including integration with geolocation data, grant and funding awards, corporate financing, and demographic data such as nationality, gender, ethnicity, or socioeconomic status among research participates
3. after sharing the data, we will document accessibility options for all kinds of people, including citizen researchers, to use it. While the primary initial userbase will be people who already use ClinicalTrials.gov, we seek to make this data accessible and interesting to undergraduate students of all disciplines and pilot data accessibility in non-English languages including Hindi, Bengali, and Swahili languages.

Open practices

Openness and FAIR data integration is a strength of this proposal which we take for granted as superior, and instead our focus regarding open data practices will be in good reporting of what content we share and documenting our publication process as a model

for others to emulate. Our publication venue is in Wikidata which has been the most popular, FAIR, and open cross-domain data repository in the world since at least 2015.

This project starts with semi-structured open data in ClinicalTrials.gov which we will map to Wikidata, thereby making it highly structured, FAIR, and accessible in the Semantic Web and in multiple languages. Perhaps more significant than our making this data FAIR is our intent to document our process as a case study to demonstrate how the data was inaccessible and not FAIR before. Currently, many researchers see ClinicalTrials.gov to be FAIR and open because they compare it to conventional data management. This project will demonstrate how much more open this data can be and what networked integration can accomplish.

Monitoring

Monitoring is a strength of this proposal which we take for granted as superior, and instead our focus regarding monitoring will be in good reporting of what monitoring we accomplish and documenting our monitoring process as a model for others to emulate.

This project will publish its output into Wikidata, the structured data general reference repository which is part of the Wikimedia platform. Since its inception in 2001, Wikipedia and the Wikimedia platform have developed a culture and community where anyone can edit and a mix of humans, human-operated semi-automated tools, and automated bots monitor the billions of edits to millions of publications which hundreds of thousands of people make every day in hundreds of languages. Our view is that in comparison to any other general interest data curation project, Wikidata provides the most openness and transparency to scrutiny and natively provides the most information about how it processes its data collections. The best way to describe the monitoring plan for this project is to say that we will use the native Wikimedia platform monitoring suite of tools and products to collect and report metrics including count of edits; count of reported changes or conflicts; count of errors identified in the source dataset; count of comments; count of active reviewers and volunteer participant editors; and audience communication impact. This project has the strength of having a designed monitoring system in place which we will not change. Instead, we will make a model report collecting the metrics which are relevant to this project, and we will document how we collect those metrics from the Wikidata platform and how we interpret them, and we will create documentation for anyone else to post data for research production into Wikidata and monitor their own projects after our model.

Success indicators include the following:

1. integration of records from 80% of ClinicalTrials.gov trials into Wikidata, with each trial having an average of 10 structured data statements of fact
2. translation of a limited vocabulary for queries and the web interface to make this data accessible in English, Hindi, Bengali, and Swahili
3. published comments - endorsement or other feedback - from a diverse community of 100 Wikimedia editors

4. publication of documentation for anyone to model this project and in advocacy of FAIR and open data

What is your plan for managing project deliverables?

The short explanation of our output management plan is that we will publish everything into the Wikimedia platform, and deposit copies into our university institutional repository, and additionally publish a research paper in an indexed academic journal. All data from this project will have a Creative Commons Zero (CC0) dedication, and all other media will have a Creative Commons Attribution 4.0 International (CC BY 4.0) license. We intend for every part of this project to be open, FAIR, and accessible to a diverse audience of users.

There are two kinds of research outputs for this proposal: structured data and prose documentation. Our primary venue for publishing structured data is Wikidata, because there it becomes available for production or export in the Semantic Web and keeps metadata labels in multiple languages for its provenance and open licensing. We will additionally publish a copy of our data as its own dataset and media product, including in Zenodo and our university's institutional repository. We will set up a Wikimedia research project page, as is customary for any project in the Wikimedia platform, and either host or link out to all media projects from that page in the established way for this platform. This research page will be a multimedia interface for accessing the data, using the data, and browsing prose documentation.

Prose documentation will include instructions for using this data in Wikidata or exporting it for use in any other context. We will also publish information about the project to encourage broad social discourse in diverse academic fields about public access to clinical trials research. Audiences which we imagine include researchers, non-research professionals, and citizens with interest in clinical research, health care, public policy, corporate finance, research funding, public health, and social disparities. While our most detailed prose explanations will be in English language, as the structured data for this project is multilingual, we will also publish our portal interface in English, Hindi, Bengali, and Swahili.

Provide an explanation of the budget

This project has two salaries - one for a researcher to oversee integration of the content into Wikidata, and another for a data scientist to provide assistance with refining the ClinicalTrials.gov data to the Wikidata model.

Student researchers will conduct evaluation of data quality, characterization of this project's dataset within the Open Semantic Web, and critique the proposal for the ethical considerations which it raises. There are two research projects planned - one for a group of data science students whose analysis will include machine learning and probably include entity disambiguation and matching. In the summer project, the students will collaborate

with the Center for Ethics and Justice at the university to document the data's accessibility and utility.

The translations will happen fairly early in the project, as they concern the labels and query terms for accessing the data, and usually not the data itself. This project begins with a substantial structured data corpus of the target languages and will build from that rather than originating any new system.

Provide budget details

- Wikidata engagement: 6 weeks FTE
- data science: 6 weeks FTE
- equipment: 0
- graduate student research: 8 weeks FTE, university rates
- undergraduate student research: 8 weeks FTE, university rates
- translation: 4 weeks FTE, market rates through vendors
- other consulting: 4 weeks FTE, market rates through vendors

Project duration

18 months

Funding program

This project was funded by Wellcome Trust via the Open Research Fund 2019.

Grant title

FAIR and open multilingual clinical trials in Wikidata and Wikipedia

Hosting institution

School of Data Science, University of Virginia

References

- Azzam A, Bresler D, Leon A, Maggio L, Whitaker E, Heilman J, Orlowitz J, Swisher V, Rasberry L, Otoide K, Trotter F, Ross W, McCue J (2016) Why Medical Schools Should Embrace Wikipedia. Academic Medicine (1). https://doi.org/10.1097/ACM.0000000000001381

- 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 https://doi.org/10.3897/rio.5.e35820
- Rawat C, Sarkar A, Singh S, Alvarado R, Rasberry L (2019) Automatic Detection of Online Abuse and Analysis of Problematic Users in Wikipedia. 2019 Systems and Information Engineering Design Symposium (SIEDS) https://doi.org/10.1109/sieds.2019.8735592
- Weiner SS, Horbacewicz J, Rasberry L, Bensinger-Brody Y (2019) Improving the Quality of Consumer Health Information on Wikipedia: Case Series. Journal of Medical Internet Research 21 (3). https://doi.org/10.2196/12450
- Wexelbaum R, Herzog K, Rasberry L (2015) Queering Wikipedia. In: Wexelbaum R (Ed.) Queers Online - LGBT Digital Practices in Libraries, Archives, and Museums. pp. 61–80. Litwin Books, Sacramento, California. [ISBN 978-1936117796].