Reliable sources and public policy issues: analysing the role of organizations as sources on Wikipedia and Wikidata

Dr Amanda Lawrence
RMIT University

Computer scientist (TBC)
RMIT University

Abstract

Government departments and agencies, civil society organizations, think tanks, research centres and consultants are prolific publishers of a range of genres including research reports, policy briefs, fact sheets and datasets which play a critical role in the circulation of research and ideas on public policy and public interest issues, yet their role in the knowledge ecosystem is often overlooked (Lawrence 2018; Lawrence 2022; Sedgwick & Ross 2020). This oversight can also be seen with studies of the reference sources on Wikipedia. Although there have been various studies analyzing Wikipedia references they tend to focus on formal publications such as academic journal articles and books or news media; there has been little systematic analysis of publications produced by organizations as reliable sources on Wikipedia. Given citations are regarded as the cornerstone of Wikipedia's verifiability and credibility, this is a major oversight which deserves attention.

This research project seeks to analyse and map citations across a range of public policy and public interest topics on English Wikipedia with a particular focus on organization publications (sometimes referred to as grey literature). Using a socio-technical approach the study will employ citation analysis, network analysis and case studies to develop a detailed picture of the diverse evidence ecosystem operating around public interest topics including analysis by location, topic area, sector and genres. The citation network will be enhanced through linking with Wikidata which will enable further analysis and classification of organizations and genres and visualisations of key policy networks. The project will also involve developing strategies for editors and readers in evaluating organizations as sources using Wikipedia and Wikidata and provide recommendations for improving guidelines that better reflect the complexity of the research publishing ecosystem in the digital age.

Introduction

Digital technologies and the internet have radically reduced the cost and complexity of production, dissemination, discovery and access to research publications resulting in disruptions to traditional publication models and new opportunities for both formal and informal media practices to expand. In addition to academic journal articles and books from commercial and scholarly publishers, and journalism and news coverage from media companies, there is a vast ecosystem of organizations engaged in producing, synthesizing and publishing policy-related

research in a range of genres (Wellstead & Howlett 2022; Williams & Lewis 2021).

Public policy is a complex, dynamic, multisector and multicentric environment that relies on a diverse evidence ecosystem (Cairney et al. 2019; Davies et al. 2019) and organization-based publishing, also known as grey literature, is particularly important for public policy and public interest issues (Lawrence 2022). Organizations from the International Panel on Climate Change (IPCC) to small-scale environmental organizations, publish material to inform and influence public debate. Research and information is regularly published by government, civil society, education and commercial organizations in a range of genres including research and technical reports, conference papers, discussion papers, working papers and preprints, evaluations, briefings, reviews, case studies, factsheets, statistics and datasets. Much of this material is available online, open access, and is produced in more timely and accessible ways than academic journal articles (Lawrence 2018). They also provide diverse perspectives including from government, community, commercial, Indigenous, and professional organizations which are not available through traditional sources. At the same time organization publications require close scrutiny and critical review as they may promote vested interests or political positions of their organization. Many reports and papers from organizations lack identifiers such as DOIs or even adequate metadata and bibliographic standards and their diverse, disaggregated and dispersed nature make evaluation and tracking of these sources extremely challenging (Sedgwick & Ross 2020).

One of the cornerstones of Wikipedia is its reliance on citations from reliable sources, however little is known about the nature and extent of citations to organization-based publications on Wikipedia and other Wikimedia platforms. A key part of Wikimedia's defense system against mis and disinformation is its content and citation policies however Wikipedia's reliable sources policies are still grounded in traditional notions of the research publishing economy as primarily commercial and scholarly publishers and traditional news media. Surprisingly while the Wikipedia guidelines on reliable sources for medicine and science include information on the importance of publications from expert bodies and organizations, the general guidelines, used by most editors, only refer to the dangers of using self-published documents. Wikipedia's reliable sources policies and editing choices flow through to other platforms such as Wikidata. As Crompton (2020) points out, 'Wikidata, which was first seeded by the content from English Wikipedia infoboxes, is also biased in favour of the type of content that is already in English Wikipedia, which itself is skewed towards the typical or traditional interests…' At the same time, Wikipedia and other Wikimedia projects have had to continually monitor and defend the site from mis and disinformation, vandalism, pranks, bias, omissions, inaccuracies and other dangers inherent in such an open project. These issues have become more prominent as our life online has expanded.

This leaves the community and Wikipedia editors facing considerable issues in terms of verifying and using organization publications, an issue replicated across the wider scholarly communication system. As the WMF White paper on Knowledge integrity notes: 'Technology platforms across the web are looking at Wikipedia as the neutral arbiter of information, but as Wikimedia aspires to extend its scope and scale, the possibility that parties with special interests will manipulate content, or bias to go undetected, becomes material.' (Zia et al 2019).

To address this important but overlooked challenge for knowledge integrity and reliable sources on Wikipedia this research project will focus on two core questions:

- What is the nature and extent of the sources cited for public policy related issues on English Wikipedia, including organization publications?
- How can policy reports and papers from organizations (grey literature) be verified and cited more efficiently and effectively on Wikipedia?

The project aims to:

- Provide new insights on what is cited on Wikipedia across social, politics, environmental, public health and policy related articles.
- Increase the transparency of vested interests on public interest issues.
- Improve recognition and coverage of diverse and credible perspectives such as the views of Indigenous, environmental and community organizations.
- Strengthen and streamline Wikipedia's citation and verifiability processes for editors and readers.
- Improve citation template infrastructure
- Contribute relevant research findings to Wikimedia projects including Wikicite, Shared citations and other projects.
- Provide an open linked dataset on organization publications for ongoing analysis and reuse
- Provide guidelines and a dataset for the wider evidence and policy community
- Enhance Wikimedia's role as a leader in digital and media literacy and education – helping to deliver the 2030 Movement Strategy as essential infrastructure of the free knowledge ecosystem.

Date: The project will be conducted part-time 2 days/week over 8 months and will run from July 2023 until 30 March 2024 including a break for Christmas and summer holidays in Australia.

Related work

Despite the important role of sources on Wikipedia they are generally understudied, certainly in comparison to wider scholarly communication and evidence-based policy topics. A large factor in this is their inaccessibility for large-scale data analyses (Singh et al 2021). There is still no standard format for references on Wikipedia and no central database of referenced sources – although this is the aim of the proposal for a Shared citations database. Generally researchers have to extract references from a data dump of Wikipedia via multiple templates and then try to classify them. Despite these challenges there have been various studies of Wikipedia sources and Arroyo-Machado et al. (2022) list 15 key publications dating back to 2007 although only one is focussed on the Humanities.

A key approach in classifying citations used has been to analyse identifiers such as DOIs or ISBNs, an approach used by WMF researchers in 2018 to stimulate research into sources on Wikipedia (WMF 2018). However this has limited value given most sources, including most policy reports and papers, or even news media do not have identifiers. As Singh et al. (2021) found in their large-scale analysis of 29 million Wikipedia citations based on identifiers, only 7% of Wikipedia pages cite a journal article with a DOI and 13% cite an item with an ISBN. The rest (80%) were described as web links. Singh et al. have made their dataset available in Zenodo and this data has been further analysed to map science (Yang & Colavizza 2022a) and humanities academic sources (Torres-Salinas et al. 2019). There has also been a study of news media sources from the same data (Yang & Colavizza 2022b), which were estimated to be

30% of citations – leaving half of WP citations still unaccounted for.
In 2022 Lewoniewskia and colleagues conducted more detailed analysis on the most used references by source across all language Wikipedias. As expected they found that a significant portion of references are to academic publications and news media however they also identified a large number of official data sources (census data) and major organizations eg WHO and UNESCO as key sources (Lewoniewskia et al. 2022). Limited access to this data is available via the BestRef website which also allows for specific urls to be searched showing various ranking metrics.

Beyond these large-scale analyses other important research has focused on specific pages, topics, or a random selection of pages (see for example Avieson 2019; 2021; Dehdarirad et al. 2018; Luyt 2021; Ford 2013). These kinds of case studies and targeted approaches provide additional insights into the vast Wikipedia citation data. These content-analysis methods are a valuable complement to more quantitative approaches as they provide a more detailed picture of the nature of the relationship between topics and citations, something particularly important for the study of organization-based publications as sources in public policy related content.

This research will also build on and extend the Missing Link Project, funded by a WMF Alliance grant in 2022 which has supported the inclusion of Australian policy reports and reputable organizations from Analysis & Policy Observatory on Wikidata. This project has engaged with the Reliable sources noticeboard to discuss specific organizations as well as the general guidelines for policy reports. This work was presented by Amanda Lawrence and Brigid Van Wanrooy at the Worlds of Wiki conference at the University of Sydney in November 2022 and a case study will be published later this year as a journal article.

Methods

To answer the two main research questions listed above the project will take a sociotechnical approach to the research methods and analytical tools including content analysis, citation and network analysis, data linking, visualizations and case studies.

The focus of analysis will be on around 1000 public policy related articles and their citations on English Wikipedia combined with data from entities on Wikidata including concepts, organizations and publishers, locations and other data. Various administrative pages on English WP will also be analysed for guidelines and policies and a number of case studies developed on key topics and organizations.

To define the public policy domain, which crosses both science and social sciences, we will start with a number of key articles and use the internal link structure of Wikipedia combined with categories, Wikidata concepts, etc. to develop a list of key topics across the public policy domain. For example based on What links here link count the *Public policy* article on Wikipedia has 2,147 direct links from other articles while the *Science policy* article has 353 and *Environmental policy* 651. Many of these policy topics have lists and portals, country specific subpages etc. which will also be analysed to provide a corpus of around 1000 public policy related articles. Consultation on the list of articles for analysis will also occur with Wiki projects such as the science policy project and some of the environmental, public health and medicine projects as well as other special interest groups.

Following the selection of content, references will be extracted and classified then mapped to Wikidata entries, topics, and locations.
The citations from the policy arena can then be compared to the full citation data for English Wikipedia. As discussed earlier, access to WP citations is not easy however there are various methods and tools which have been developed by other researchers which are available as well as existing datasets of citations.
Arroyo-Machado et al. (2022) provide a summary table of Wikipedia data sources by format, update frequency, data quantity, type, and challenges which includes: Wikimedia Dumps, MediaWiki and Wikimedia APIs, Wiki Replicas, Event Streams, Analytics dumps, WikiStats, Dbpedia, XTools, Repositories and Altmetric aggregators. It is expected that for this research Wikipedia data dumps, web scraping from the target pages, and citation data sets from previous research will be the main data source for citations. These will be linked and enhanced with data from other databases such as Wikidata, CrossRef, ISNI, OpenAlex, Dimensions, Internet Archive etc.

The final dataset will then be analysed for frequency of citation, type of organization, and visualized using various tools such as network graphs, timelines, geospatial mapping etc.
A rating of the reputation of sources will be made based on the information available on organizations via WP and WD, the reputable sources lists and other sources and where poor sources have been listed these may be flagged on the relevant pages. The data extraction, linking and analysis process will be assisted by a data scientist working on the project for 2 months.

Consultations and feedback with the Wikimedia community will occur at Wikimania in Singapore in August and the Wikidata conference in Taiwan in September 2023 and online with various projects including Wikicite and the Shared citations project. Funding for attending the Wikidata Conference in Taiwan is included in the budget.

Following an analysis of guidelines available on WP and consultation with various projects such as science and medicine and other interest groups a set of draft guidelines for grey literature will be developed and circulated and the data, a project report and journal article will be published open access.

Timeline

2023

July

- Project initiation, literature review, analysis of existing data, initial page selection and analysis
- Set up project page on Meta and Github or OSF
- Engage data analyst to the project for Aug-Oct

August – September

- Consultations with community at Wikimania, Singapore in August
- Review policy pages in WP and organizations in Wikidata and make selection of corpus
- Extract citation data from corpus into structured format
- Data cleaning and linking

October

- Data cleaning and linking
- Consultation with community at Wikidata Taiwan on data analysis

November - December

- Data analysis and visualisation

2024

Jan (summer break)

February

- Case study analysis, data synthesis and initial results write up

March

- Guidelines developed and circulated for review and feedback including giving presentations to key groups across the Wikimedia community
- Report and journal article drafts prepared.

April

- Publication of report, data and methods in open access report on Zenodo, OSF or Github
- Journal article submission
- Project completion and reporting.
- Guideline development ongoing with community and proposal for changes

Expected output

Research progress will be documented on a project page on Meta and more detailed information maintained on Github or Open Science Framework (OSF) or other suitable location. Engagement with the community will occur as the project progresses via presentations and meetings online and in person at Wikimania in Singapore and Wikidata Taiwan. Other travel for in person meetings may occur if the opportunity arises.

Results will be disseminated via a project report published on Zenodo or other suitable database, and through an academic open access journal article and conference papers. A project report will be published on Meta and various presentations will be given to the Wiki community and the wider public policy and research community.

Datasets of citations and organizations will be published on Zenodo under a CC BY-SA 4.0 license and where possible Wikidata will be used to collate and expand data on organizations.

Interactive visualisations will be part of the topic modelling process and the citation modelling using various open source tools such as Wikidata's Scholia, Networkx (https://networkx.org/), a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks, or Voxviewer, a software tool for constructing and visualizing bibliometric networks, or other tools as appropriate.

Risks

In the original EOI there was only one researcher, making the project vulnerable if the lead researcher needed to step out for any period of time. This has been mitigated by the inclusion of another project member who will also provide an expanded set of skills required for this project.
Otherwise there are few risks to this project.

Community impact plan

The need to improve references from diverse sources is well recognised by the Wikimedia community, as is the need for improved processes and system architecture for citations. The Wikicite community will be a key group for engagement on this project and as it has recently begun meeting again after a break over the last few years this is a perfect time for this project. I have previously been involved in events and conferences on Wikicite in Australia and at the 2020 online conference. There is strong interest in citations and Wikidata in the Australian community and I will draw on the expertise of community members such as Alex Lum, Thomas Shafee, Bunty Avieson and Heather Ford as well as my colleagues at RMIT University.

The project will also provide data and case studies for the Shared Citations proposal which I am following closely. As with the Shared

Citations project, this research aims to make citations easier for editors and more useful for readers and more efficient for the Wikimedia architecture across projects.

As an active member of Wikimedia Australia and current President of the Committee, I am deeply engaged in the Australian community as well as active across ESEAP. I am organising a Wikimedia research roundtable in July at the University of Sydney where I will be able to present the project and engage with other researchers.

Consultations with the wider community will be able to occur in person at Wikimania in August 2023, and the Wikidata Conference in Taiwan in October 2023 as well as online via presentations and webinars in early 2024 when results are available.

There are a number of relevant WikiProjects on English Wikipedia that may be interested in this research including: WikiProject academic journals which provides statistics on citations using the {{cite xxxx}} template including {{cite report}} (3,139); Science; Science Policy; Politics; Climate Change etc. Engaging across projects and community interests will be a key aim of the project as we develop guidelines and recommendations.

At a broader level, the project supports the 2030 Movement strategy in four key areas:

- Improve user experience: by supporting editors working on public policy issues to identify reliable sources from a wider range of reputable sources;
- Manage internal knowledge: improving guidelines and processes for dealing with organizations as reliable sources for social and policy issues;
- Identify topics for impact: supporting the editing process on key public policy issues such as climate change, social inclusion and public health;
- Innovate in free knowledge: improving the way organization publications are managed making them easier to find and evaluate for the wider community.

Evaluation

The project will be evaluated on the basis that:

- A corpus of 1000 articles and their citation data will be extracted, linked and enhanced, visualised, and results made public for reuse
- The research methods will be made available along with the dataset for replication across other Wiki projects.
- Guidelines will be drafted and shared with the WP community.
- Wikidata will be significantly expanded with information on key organizations.
- A report will be published for both the Wikimedia and wider evidence and research community to share the findings and improve understanding of organization publishing in public policy

Budget

The main expenditure is for

- a part-time salary for Dr Amanda Lawrence estimated at 2 days/week for 8 months based on an Australian academic salary level B = AUD38,848/US$25,964

This funding will provide dedicated time and resources for Lawrence to work on this project based at RMIT University. Working part-time will allow time for community engagement and feedback over a longer period.

- a part-time salary for 2 months at 2 days/week for a casual data analyst to assist with data extraction, linking and visualization= AUD8363/US5588
- Travel and accommodation to attend the Wikidata conference in Taiwan in

September 2023 to consult with researchers and the Wikimedia community and share findings = AUD$5000/US3,337
- Publication costs, software and miscellaneous expenses= $2000
- Organisation overheads of 15% of the project budget as per WMF guidelines.

Total budget: AUD62,342/USD42,146

References

Amaral, G et al. 2021, 'Assessing the Quality of Sources in Wikidata Across Languages: A Hybrid Approach', *Journal of Data and Information Quality*, vol. 13, no. 4, p. 23:1-23:35.

Arroyo-Machado, W, Torres-Salinas, D, & Costas, R 2022, 'Wikinformetrics: Construction and description of an open Wikipedia knowledge graph data set for informetric purposes', *Quantitative Science Studies*, vol. 3, no. 4, pp. 931–952.

Avieson, B 2019, 'Breaking news on Wikipedia: collaborating, collating and competing', *First Monday*, vol. 24, no. 5, viewed 13 March 2020, <https://firstmonday.org/ojs/index.php/fm/article/view/9530>.

Avieson, B 2022, 'Editors, sources and the "go back" button: Wikipedia's framework for beating misinformation', *First Monday*, viewed 9 November 2022, <https://firstmonday.org/ojs/index.php/fm/article/view/12754>.

Cairney, P, Heikkila, T, & Wood, M 2019, *Making policy in a complex world*, Cambridge University Press, Cambridge, UK.

Crompton, C et al. 2020, 'Familiar Wikidata: The Case for Building a Data Source We Can Trust', *Pop! Public. Open. Participatory*, , no. 2, viewed 23 January 2021, <https://www.popjournal.ca/issue02/crompton>

Davies, H et al. 2019,'Conclusion: lessons from the past, prospects for future', in A Boaz et al. (eds.), *What works now? Evidence-informed policy and practice*, Policy Press, Bristol, UK, pp.369–382.

Dehdarirad, T, Didegah, F, & Sotudeh, H 2018, 'Which Type of Research is Cited More Often in Wikipedia? A Case Study of PubMed Research', viewed 7 December 2022, <https://hdl.handle.net/1887/65240>.

Ford, H et al. 2013,'Getting to the source', in, *Proceedings of the 9th International Symposium on Open Collaboration*, viewed 13 September 2022, <http://dl.acm.org/doi/10.1145/2491055.2491064>.

Kopf, S 2022, *A Discursive Perspective on Wikipedia: More than an Encyclopaedia?*, Springer International Publishing, Cham, viewed 14 December 2022, <https://link.springer.com/10.1007/978-3-031-11024-5>.

Lawrence, A 2018, 'Influence seekers: the production of grey literature for policy and practice', *Information services and use*, vol. 37, pp. 389–403.

Lawrence, A 2022, 'Research use and publishing diversity: The role of organisation research publishing for policy and practice', *Australian Journal of Public Administration*, vol. 82, no. 1, pp. 46–68.

Lewoniewski, W 2022, 'Identification of Important Web Sources of Information on Wikipedia across various Topics and Languages', *Procedia Computer Science*, vol. 207, pp. 3290–3299.

Lewoniewski, W, Węcel, K, & Abramowicz, W 2020, 'Modeling Popularity and Reliability of Sources in Multilingual Wikipedia', *Information*, vol. 11, no. 5, p. 263.

Priem, J, Piwowar, H, & Orr, R 2022, 'OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts', viewed 28 March 2023, <http://arxiv.org/abs/2205.01833>.

Sedgwick, RE & Ross, R 2020, 'Making Grey Literature Discoverable and Impactful on JSTOR Through Comprehensive Search and Rich Metadata', *The Serials Librarian*, vol. 79, no. 3–4, pp. 261–266.

Singh, H, West, R, & Colavizza, G 2021, 'Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia', *Quantitative Science Studies*, vol. 2, no. 1, pp. 1–19.
Thorpe, K, Sentance, N, & Booker, L 2023, *Wikimedia Australia and First Nations Metadata: ATSILIRN Protocols for Description and Access*, University of Technology Sydney, viewed 31 March 2023, <https://opus.lib.uts.edu.au/handle/10453/166700>.
Torres-Salinas, D, Romero-Frías, E, & Arroyo-Machado, W 2019, 'Mapping the backbone of the Humanities through the eyes of Wikipedia', *Journal of Informetrics*, vol. 13, no. 3, pp. 793–803.
Wellstead, AM & Howlett, M 2022, '(Re)Thinking think tanks in the age of policy labs: The rise of knowledge-based policy influence organisations', *Australian Journal of Public Administration*, vol. 81, no. 1, pp. 224–232.
Williams, K & Lewis, JM 2021, 'Understanding, measuring, and encouraging public policy research impact', *Australian Journal of Public Administration*, vol. 80, no. 3, pp. 554–564.
WMF 2023, *Wikicite/Shared citations*, Wikimedia Foundation.
Wong, K, Redi, M, & Saez-Trumper, D 2021,'Wiki-Reliability: A Large Scale Dataset for Content Reliability on Wikipedia', in, *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.2437–2442, viewed 7 December 2022, <http://arxiv.org/abs/2105.04117>.
Yang, P & Colavizza, G 2022a,'A map of science in Wikipedia', in, *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France. A*, <https://wikiworkshop.org/2022/papers/WikiWorkshop2022_paper_4.pdf>.
Yang, P & Colavizza, G 2022b, 'Polarization and reliability of news sources in Wikipedia', viewed 29 March 2023, <http://arxiv.org/abs/2210.16065>.
Zia, L et al. 2019, 'Knowledge Integrity - Wikimedia Research 2030', viewed 4 December 2022, <https://figshare.com/articles/journal_contribution/Knowledge_Integrity_-_Wikimedia_Research_2030/7704626/2>.