Extended: Wikimedia and Digital Archives India: Investigating Citational Practices and Backlinking as a Strategy to Increase Discoverability and Stakeholders in the Movement.

Puthiya Purayil Sneha
Independent Researcher

Dr. Soni Wadhwa
SRM University, Andhra Pradesh

Abstract

In this research we propose to examine the current state of citational practices on Wikimedia projects, paying specific attention to resources available in GLAM institutions (Galleries, Libraries, Archives, and Museums), with a focus on Indian languages. Coordinating across stakeholders is one of the focus points for the Wikimedia Movement as stated by the 2030 Wikimedia Strategy, where the mandate is to explore different ways of working with people in the movement but also with external stakeholders who align with the vision of contributing to the digital knowledge commons. GLAM institutions are a key contributor to this vision, given their collections of cultural content spanning languages, forms and formats. Therefore, in this study we explore how GLAM collections, specifically Indian language archives, are cited and reused across various Wikimedia projects, and in what ways this synergy between the two can be enhanced. We will undertake a mixed methods study, firstly to understand the prevalence and use of GLAM content across Wikimedia projects, with a focus on the practices of citation and backlinking. Second, we aim to map existing collections of archival content across specific Indian languages, to understand some of the challenges in digitising and bringing these collections online. Finally we aim to develop and offer recommendations towards a stronger WikiGLAM relationship and strategy for India.

Introduction

The nature of dissemination and consumption of information online has led to debate around the ways in which seekers of information often end up accessing certain kinds of content on the internet.(Dutton et al. 2017) In recent years, with increasing algorithmic curation of content and evolving forms of mis/disinformation online, the impact of these developments on belief making while reinforcing stereotypes, and potential adverse consequences for our political and social lives has also been studied.(Lazer, Swire-Thompson, and Wilson 2024) One of the causes of this phenomenon of reinforcing existing beliefs is a lack of easy access to diverse primary sources for citation, in the public domain or via open knowledge platforms like Wikimedia projects. GLAM institutions, especially archives, are home to primary sources that can offer robust references to help with diversity and veracity of information online. Further, these sources are critical for the ways in which information disorders (Wardle and Derakhshan 2017) can be addressed. Wikimedia projects have the mandate of addressing issues around information, and specific projects like Wikisource as well as Wikipedia are going to play an important role in this mandate.

The digital turn has led to a significant change in the nature and scope of cultural

archives, spanning the process of their creation, management and use. Along with opportunities for advancement in methods of preservation, storage, curation and dissemination there are also several challenges in working with digitized, or born-digital content. Especially in the Indian context, given specific concerns related to access, infrastructure and linguistic barriers, the process of digitalization of archives is a complicated one. With the rise of more niche and private archival spaces today, questions of privacy, ownership and access have also become paramount. A lot of new archival initiatives are focused on developing accessible, networked and dynamic digital collections that are able to address the above challenges. However archival content also faces a problem of discoverability and use, for a range of reasons from lack of good linked, structured metadata to challenges specific to certain languages and new forms of multimodal content. Collaborations with open knowledge platforms like Wikimedia projects can go a long way in improving the discoverability and use of archival collections, and simultaneously enhance the quality and expanse of content on Wikimedia projects as well.

Given the diversity and number of languages in this region, South Asia Wikimedia projects need closer attention in terms of content quality and quantity, and community building. Research on Indian language Wikimedia projects also shows the possibilities of work that can be undertaken in terms of diversifying the nature of content and the quality of sources. For instance, the English Wikipedia entry on Mahatma Gandhi links to The Collected Works of Mahatma Gandhi in 100 volumes on Wikisource (where some volumes have been made available by the community). However, there is an independent archive called Gandhi Heritage Portal which hosts all these 100 volumes. This archive portal is not linked to Mahatma Gandhi's Wikipedia article page. This page reports that around 500 other Wikipedia article pages link to it. This information about how "networked" this page is can be enhanced by listing non-Wikimedia pages that it links to. A cursory glance at the references section on the page reveals that most sources are secondary sources on Gandhi. Some sources are also from the news media. However, the fact that the most primary source of information (the collected works of Gandhi) is not cited implies that information about Gandhi is gleaned from sources that do not come from his writings. We take this as a cue to ask the following research questions:

- What is the Wikimedia movement's philosophy around primary and secondary sources? What does it say about its information ecosystem? Given the focus on the (evolving) nature of online information, how can it undertake a critical approach towards facilitating access and use of primary sources?
- What is the synergy between online archival collections and Wikimedia projects in Indian languages? What are the citational and backlinking practices currently being used?
- What are current challenges and affordances in the digitalisation and open access to archival collections? What can strengthen the WikiGLAM strategy in working towards the goals of the open knowledge movement, and in battling information disorders and facilitating better access to data?

It is hoped that these questions will trigger a larger conversation about the ways in which the Wikimedia movement, and the open knowledge movement at large can play a role in making primary sources of information more widely accessible, and fostering discoverability and reuse. The focus of the current study is on India,

namely Indian languages and topics pertaining to India, including projects and pages in English. While focused on India, the insights from the study are likely to be important for large parts of the majority world where regions share certain features: low resource languages, history of colonization, unequal access to information, and challenges pertaining to community building, technical training, and opportunities for collaboration across Wikimedia projects. The outcomes – such as research findings, strategy recommendations for WikiGLAM, a dashboard for tracking citation to and from Wikimedia, a mapping of Indian language archival collections, and community consultations– will advance the Wikimedia movement's strategy and objectives of becoming an essential infrastructure for free and open knowledge. This will be a two year, extended research project, **beginning July 1, 2025 and ending on June 30,2027.**

Related work

The research questions emerge from earlier work undertaken by the researchers in this proposed project, in the areas of open knowledge, GLAM and digital archival practice, and Indian languages, as part of the Access to Knowledge team and other programmes at the Centre for Internet and Society. These include short term research studies that mapped GLAM institutions in a particular region, and data gaps and challenges related to content creation on specific Wikimedia projects and Indian languages. (Access to Knowledge Programme, 2021) A recent report on the state of the open movement in India further identifies some of the changes in the understanding of 'openness' over the last decade, and emerging areas of intervention. (Wadhwa 2024) A report mapping the state of digital humanities, (Puthiya Purayil 2016) and continuing work on the digital turn in archival practice in India (Puthiya Purayil 2022) also foregrounded some of the challenges and affordances of digitalisation in knowledge infrastructures, with a focus on shifts in digital archival practice and the possibilities for better access to cultural resources. Most recently, there has also been an effort to collaborate with other key stakeholders to explore the intersections of AI, archives and Indian languages, and their implications for the future of the global knowledge commons.(Wadhwa et al 2024)

A growing body of research on citational politics (The Citational Justice Collective et al. 2022), although focussed primarily on academia, has highlighted some of the concerns we wish to explore further in this project - namely around the visibility and discoverability of knowledge from underrepresented regions and languages across the majority world. Key technological and infrastructural gaps here include that of digitisation and visibility of content, and how this impacts the search and retrieval of information online. Wikimedia projects, as a form of knowledge infrastructure, also rely heavily on practices of citation, and it would be important to examine how these practices have transformed on a global, open knowledge platform. These would offer insights on discoverability and use of content, and foster engagement with some of these challenges around open access to cultural content.

Methods

The above questions on the role archives can play in the Wikimedia movement and how Wikimedia can advance the goals of archives require a mix of qualitative and quantitative analyses. Different aspects of this study will therefore engage with specific methods. These are listed below.

Desk Research, Literature Review and Sampling: The first phase of the study will comprise a short period of desk research and literature review, to identify specific Indian language Wikimedia projects, thematic areas and topics of focus for this project.Given the focus on archives and GLAM more broadly, the project would be located within the areas of arts and culture, literature and heritage. We will also undertake a detailed literature review of current research on the digital knowledge commons and the impact of mis/disinformation, algorithmic curation, including the growth of LLMs and generative AI on content creation, discoverability and use of open knowledge resources. Based on the above, we will arrive at a defined sample of pages on 2-3 selected Wikipedia projects for study.

Data collection: Data collection will be undertaken in two ways:

1. **Web scraping**: Selected Indian language Wikimedia projects will be scraped with the help of a relevant tool to check for:
    - citation or references to primary sources
    - nature of back linking on wiki projects and
    - links incoming to wiki projects

Based on the picture of backlinking that emerges from the projects, further inquiry will engage with people in the ecosystem of free knowledge.

2. **Interviews and Focus Group Discussions**: We aim to conduct at least 22 interviews and 3 focus group discussions with:
    - Stakeholders at various archival institutions
    - Wiki contributors and editors
    - Subject experts, including but not limited to academics, technologists, legal professionals, and creative practitioners working in the GLAM sector.

This will help us survey the nature of citational practices on Wikimedia projects and help understand:

- The nature of training of Wikimedia community/contributors regarding sources
- Subjective choices involved in citation processes
- Topics and domains that can be enhanced with citations of archives

Further, it will also help us collect insights on the kind of archival collections in Indian languages currently available online, what are challenges in digitalisation and open access to their content, and how they imagine the (re)use of their collections on the internet.

Analysis and Writing

The data collected through the above approaches will undergo both quantitative and qualitative data analysis and coding, to identify key themes and patterns. The key learnings will be compiled in the form of a final report and recommendations for a WikiGLAM strategy for India, and the analysis will further be used to develop the other outputs from this study. For instance, data visualization of key learnings/patterns from the quantitative data will inform the creation of the interface of the planned dashboard (discussed below). The data and its quantitative analysis is also likely to be published as a dataset on Wikidata (discussed below). The insights from the qualitative data will also be shared as part of research publications.

Expected output

The findings of the study will be made available for the community and researchers in the open movement in the following ways.

1. **Dataset for Wikidata.** What Wikimedia projects link to offers critical insights in terms of a dataset that will promote introspection across all projects regarding citation practices. This can contribute to the work on building structured and linked open data, and also help the community strategise training for better citation and fact-checking processes.
2. **API and Dashboard**: A dashboard for tracking and measuring links to and from archival collections and Wikimedia projects will be created. While publicly accessible, the key stakeholders here are community members and decision makers on Wikimedia projects, and GLAM institutions who can monitor quality and quantity of citations and identify topics for impact.
3. **WikiGLAM strategy**: Recommendations based on findings will be shared with Wikimedia leadership, community, and researchers in the form of a final report aimed at developing a robust WikiGLAM strategy for Indian languages. These will comprise recommendations on capacity-building, community engagement, content creation and research on integrating archival collections with Wikimedia projects.
4. **Collaboration and Partnerships:** The Wikimedia movement seeks to partner with those who share its mission to work towards open knowledge. The learnings from the study will identify ways of liasoning with archival institutions working in Indian languages and develop roadmaps to work towards access to open resources in the long run.
5. **Resources for Wikisource**. Wikisource is emerging as an important project in making primary resources available for easy access. Working with archival institutions can help identify more resources while also offering platforms to archives to showcase their work, while also leveraging their skills and capacity building to further strengthen the Wikisource ecosystem.
6. **Advancing AI and Mis/disinformation Research:** The findings from the study are likely to offer some early insights into the prevalence of algorithmic curation and use of LLMs and generative AI in cultural work in Indian languages, and the challenges posed by mis/disinformation. These will also be shared as part of the final report and inform the strategy recommendations.
7. **Research Publications:** We will aim to circulate findings from the study among all key stakeholders outside of the Wikimedia movement as well, especially researchers and professionals working in the GLAM sector to build on this work and study context-specific regional factors. This would be through single, and co-authored research publications across diverse academic/non-academic forums. The findings will also help different communities working in the areas of arts and culture, and heritage, strategise and reflect on their engagement with the open knowledge movement.
8. **Events**: Consultations will be conducted at the beginning and towards the end of the project to bring together key stakeholders and understand the processes and workflows among Wikimedia communities and archival institutions, and disseminate learnings. These forums will also bring the two communities together to facilitate better understanding of these topics among them.

Risks

A limitation with the study comprises potential challenges with access to key stakeholders and engagement with the communities on questions related to citational politics. In the case of the former, we aim to use snowball sampling as a means to find relevant stakeholder participants in the field, in addition to our own sampling and recruitment methods. The discourse on citational politics is still relatively new in India, and is informed by research and practice emerging from non-dominant languages and communities, including feminist work. It would be important to understand how a large, open knowledge platform is able to engage with questions and challenges around citation, discoverability and use of primary archival collections online, and the broader question around what kinds of information is more easily accessible on the internet.

Community impact plan

1. **Training in Citation**. With the findings, API/dashboard, and outreach events the Wikimedia communities will have a number of spaces to meet, examine and upgrade its citation practices.
2. **Training in APIs that monitor backlinks to and from Wikipedia**. Potential for training around monitoring quality and quantity of citations will help the community strategise its relationship with archives, thereby broadening its understanding of sources of information.
3. **Access to information for users in majority worlds**. The India-specific study is likely to help Wikimedia projects from other regions of the majority world in terms of identifying resources for citing information. Additionally, it is also likely to provide a roadmap for projects in non-dominant languages to undertake their own context, region, discipline, and language specific studies.
4. **Representational parity in citational practice**. Information and related resources available online are skewed towards visible and dominant languages and communities. To identify resources for marginalised communities, the Wikimedia community needs to develop a better understanding of how to locate information about these groups. Archives related to gender, sexuality, anti-caste and feminist work for instance can facilitate this understanding and thus contribute to a richer, equity-oriented Wikimedia.
5. **Contribution to AI and data research communities**. The insights on AI and mis/disinformation in the curation of information online would also contribute to further research related to collection and management of data, especially as relevant to the majority world. Digitalisation and data management are integral processes of the work of algorithmic curation of information, and we aim to offer some insights into these as part of the study.

Evaluation

The proposed study can be evaluated on the following fronts:

1. **Examination of citation practices** on Wikipedia projects over a span of 18 months can help measure the ways in which these projects show changes qualitatively and quantitatively. Questions such as increase/decrease in citation of archival resources can be one

metric to understand the synergy between Wikimedia and archives.

2. **Further research** published on information disorders and the role of archival sources in addressing these by peers can demonstrate if the proposed initiative has been useful to a diversity of stakeholders in other related areas of work on open knowledge. This further research can be in domains as varied as intellectual property rights, coverage of primary sources on Wikisource, diversity in citational practices, information disorders vis-a-vis primary sources of information, discovery of new sources through Wikimedia pages, and so on.
3. **Greater liaising between Wikimedia projects and archival institutions**, through collaborative events and other partnerships would offer insights on the contributions of the project. It can also help Wikimedia communities determine what can be done to advance the cause of open knowledge, especially through long-term strategic partnerships.
4. **Critical analysis of archives as data** and the importance of curation and the human in the loop question are among intangible outcomes that need further engagement. One thinks of data as cleaned or organised arrays of information fed into AI models, such as LLMs. Archives can challenge that notion by showing that data/input is very diverse in nature. The skills and conceptualisation that go into the making of archives can raise the standards of data and its processing that go into the making of LLMs.
5. Whether the **interest in digitisation projects**, especially among smaller, under-resourced Wikimedia projects and archives increases as a result of working on this project, thereby addressing data inequality from a majority world perspective, is another parameter that can be used to evaluate the impact of the proposed research. This, in turn, can further advancement of technology in low resource languages thanks to interest in digitisation.

Budget

Details of the budget may be seen here:
WMF Research Fund Budget.2025

References

Access to Knowledge Programme, 2022. Research Studies on Indian Language Wikimedia Projects, Wikimedia Commons, accessed April 16, 2025. https://meta.wikimedia.org/wiki/CIS-A2K/Research

Dutton, William H., Bianca Christin Reisdorf, Elizabeth Dubois, and Grant Blank. 2017. "Social Shaping of the Politics of Internet Search and Networking: Moving Beyond Filter Bubbles, Echo Chambers, and Fake News." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2944191.

Lazer, David, Briony Swire-Thompson, and Christo Wilson. 2024. "A Normative Framework for Assessing the Information Curation Algorithms of the Internet." *Perspectives on Psychological Science* 19 (5): 749–57. https://doi.org/10.1177/17456916231186779.

Sneha, Puthiya Purayil. 2016. Mapping Digital Humanities in India. Centre

for Internet and Society. https://cis-india.org/papers/mapping-digital-humanities-in-india

Sneha, Puthiya Purayil. 2022. "Alternate Histories of Digital Humanities: Mapping the Archival Turn." In *Global Debates in Digital Humanities,* edited by Fiormonte, S. Chaudhuri and P. Ricaurte, 15-30. Minneapolis: University of Minnesota Press.https://dhdebates.gc.cuny.edu/read/global-debates-in-the-digital-humanities/section/6c852e0e-190b-4f97-991c-0472408ee4b5

Wadhwa, Soni.2024. Open Movement in India (2013-2023): The Idea and Its Expressions. Centre for Internet and Society. https://cis-india.org/a2k/blogs/open-movement-in-india-idea-and-its-expressions

Wadhwa et. al. 2024. Future of the Commons: A Conversation on Artificial Intelligence, Indian Languages, and Archives Conference Report. Centre for Internet and Society. https://cis-india.org/raw/report-on-the-future-of-the-commons

Wardle, Claire, and Hossein Derakhshan.2017. Information disorder: Toward an Interdisciplinary Framework for Research and Policymaking. Vol. 27. Strasbourg: Council of Europe. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html