Wikipedia citations in low-resource languages: challenges and opportunities

Dr. Malka Older
Global Voices/Arizona State University

Miora Radifera
Global Voices

Abstract

In this research we propose to investigate how independent news media and Wikimedia groups can collaboratively boost content and citation of small language Wikipedias. Our findings will improve the experience both for editors and readers, encouraging greater participation with and use of Wikipedia. This is also fundamental to the Wikimedia Movement's 2030 strategic direction, which prioritizes knowledge equity and "knowledge and communities that have been left out by structures of power and privilege." The research will focus on the specific case of the Malagasy language, but will include a survey of four other languages for comparison purposes, and we aim to conclude the research with recommendations for other language groups.

Introduction

Wikimedia groups working in smaller languages often struggle to find sufficient citations for their articles in the same language. The use of references in other languages stymies some users and undermines efforts towards knowledge equity.

This research will build on a joint Global Voices-Wikimedia Community User Group Madagascar edit-a-thon initiative to address the lack of citations in Malagasy by using the large corpus (over 30,000) of Global Voices articles translated into that language. Through qualitative interviews and the development of quantitative indicators and tracking, we will document the impact of that collaboration. A broader survey of the citation situation for four other Indigenous languages will allow us to consider possible approaches for other language groups.

Our main research questions include:

- What impact does the incorporation of references in Malagasy Wikipedia have on the presence and representation of that language in the Wikipedia ecosystem?
- What effect does the increased availability of references in Malagasy have on the creation of content in Wikipedia in that language?
- What impact does the incorporation of references in Malagasy have on the consumption of Wikipedia content in that language?

Date: July 1, 2025- June 30, 2026

Related work

There is little research directly in this area. Wikipedia citations themselves are non-trivial to study, and they are more often studied in terms of their usefulness as a scientific impact indicator or in reference to the value of the sources. Moreover, small-size Wikipedias of low-resource languages tend to get less research attention.

The importance of references, diversity of references, and localized references for the Wikipedia project:
"#1Lib1Ref: Calling librarians around the world to improve Wikipedia" Felix Nartey, 14 May 2021

Singh et al. 2021 note the difficulty of extracting and studying citations in Wikipedia, which they do only for the English page:
"Wikipedia Citations: A Comprehensive Data Set of Citations with Identifiers Extracted from English Wikipedia." Singh, Harshdeep, et al. Quantitative Science Studies, vol. 2, no. 1, 2021, pp. 1–19

Wikipedia @ 20: Stories of an Incomplete Revolution Joseph Reagle, Jackie Koerner, eds., 2020

"Towards a Wikipedia For and From Us All" Adele Vrana, Anasuya Sengupta, and Siko Bouterse, 2019

Methods

We propose to make this project as participatory as possible. Malagasy researchers will take the lead in investigating the impact of the collaboration on same-language citations, with support from the Global Voices research and training team. The broader, multi-country surveys will be conducted more centrally, but with the Malagasy researchers included in the design and analysis.

We will begin with a brief survey of Malagasy and four other low-resource language Wikipedia groups: Igbo, Javanese, and two others. We will sampling the percentage of same-language references and learning from editors and readers about their preferences and constraints.

The researchers, in consultation with the Global Voices research and training team, will design in-depth qualitative interviews about the processes, challenges, and possibilities of same-language citation; they will also work together on developing indicators for tracking same-language citations in different language contexts.

These indicators will be used to quantitatively compare the use of same-language citations before and after the edit-a-thon and other initiatives. Similar indicators will be used to observe the effects on readers.

A further round of qualitative interviews at the end of the project will look for changes in knowledge, practices, and attitudes among participants.

At every stage of the project, data will be discussed with Wikimedia and Global Voices communities. The interpretation of the data will be a joint project between the professional researchers and the communities, and published work will be explicitly based on those discussions.

Expected output

A toolkit for Wikipedia language groups
Essays to be published in Global Voices, in English, Malagasy, and other languages
A collaborative academic article about the project, designed as a cascading discussion among the researchers and research support team.

Risks

This is a fairly low-risk project. We will not be collecting personal or sensitive data; surveys can be anonymous. The major risks are delays due to issues such as Internet outages or possibly currency depreciation or political turmoil in any of the locations where we have staff.

Community impact plan

This project will work closely with Wikimedia affiliates, including Malagasy, Dagbani, and Igbo and Wikipedia editors. We expect that the project will lead them into thinking more carefully about their use of citations. We also hope that the collaboration between Global Voices and Wikimedia will inspire not only further cooperation between these two groups, but also greater engagement with other actors in language revitalization ecosystems.

Evaluation

How will you, and how should others, including the Research Fund chairs, evaluate your proposed project? What will you consider a measure of success for your proposed research?

- A framework for better assessing the availability and usage of same-language citations in Wikipedia posts
- A better understanding of the impact of same-language citations
- A set of clear, actionable suggestions for Wikimedia language groups seeking to increase the percentage of same-language citations.
- Serves as inspiration or guidance for other language communities

Budget

Research Fund Budget: Wikipedia citations in low-resource languages

References

Nartey, Felix "#1Lib1Ref: Calling librarians around the world to improve Wikipedia" 14 May 2021

Reagle, Joseph and Jackie Koerner, eds.,*Wikipedia @ 20: Stories of an Incomplete Revolution* MIT Press, 2020

Vrana, Adele, Anasuya Sengupta, and Siko Bouterse "Towards a WikipWikipedia @ 20: Stories of an Incomplete Revolutionedia For and From Us All" 2019

Singh, Harshdeep, et al. "Wikipedia Citations: A Comprehensive Data Set of Citations with Identifiers Extracted from English Wikipedia." Quantitative Science Studies, vol. 2, no. 1, 2021, pp. 1–19, https://doi.org/10.1162/qss_a_00105.

Halitaj, Aida, and Arkaitz Zubiaga. "Providing Citations to Support Fact-Checking: Contextualizing Detection of Sentences Needing Citation on Small Wikipedias." Natural Language Processing Journal, vol. 8, 2024, pp. 100093-, https://doi.org/10.1016/j.nlp.2024.100093.