Enhancing Discoverability, Searchability and Citability through Re-Archiving Odia-language Texts

Subhashish Panigrahi
O Foundation

Abstract

Books in scanned image form without optical character recognition (OCR) are undiscovered by readers, as in India's Odia language. Similarly, Wikipedia editors also struggle to find citable content. This issue is caused by poor infrastructure that restricts in-platform image-to-text extraction. To address this issue, over 14,000 Odia-language texts from the digital archive Odia Bibhaba were re-archived on the Internet Archive. The text was extracted and indexed on the web through in-platform OCR. Users searching for keywords can now find the full text. The texts, the content therein, and the corresponding authors have significant opportunities for Wikipedia and Wikisource.

Keywords: glam, book archiving, metadata, digitization, ocr, odia language

Background

Odia is one of India's 22 official languages, with 45 million speakers, and is primarily spoken in the Odisha state. Though provincially dominant, digital archiving of old Odia texts has been minimal. Many digital archives do not follow standard cataloguing practices, and their hosting and site configuration choices have led to the failure of web indexing.

1 State of Odia-language Public Archives

Digital public archives efforts in Odia have been meagre. Odia Bibhaba (OB, https://odiabibhaba.in) (Mishra, 2017) is arguably the largest and longest-standing digital public archive in Odia (Panigrahi, 2022). The Odia Virtual Academy (OVA) is the second-largest and the largest state-run digital public archive, complementing OB. A minuscule percentage of Odia books, magazines, and other periodicals remain available publicly in other public and private institutional portals.

The not-for-profit Srujanika began scanning and digitally archiving printed texts in 2006 and established OB in 2017. Later, the SOA Centre for Preservation, Propagation and Restoration of Ancient Culture and Heritage of India (PPRACHIN) became Srujanika's predecessor since the latter plans to retire from active archiving. Jeeban Kumar Panda, Srujanika's long-time archivist, heads PPRACHIN's archiving work. Currently, Srujanika and PPRACHIN scan and make texts available on OB (Siksha 'O' Anusandhan, 2021). By June 9, 2024, this portal hosted 13,550 book titles, excluding over a thousand periodicals (Srujanika, 2024).

2 Access Issues with Odia digital Archives

Srujanika hyperlinks OB's uploaded book files on Google Drive from sortable and searchable bibliographic data tables. Hence, books' content text is not extracted from image PDFs as Google does not integrate optical character recognition (OCR) for image PDFs, hindering web indexing. In the case of books hosted on OVA, the content text is typed but displayed as ePUBs where text search, right-click, and web indexing are disabled. Both these portals' content remains largely search-prohibitive, and each title lacks a permalink, with some bibliographic metadata being web-searchable. It is hard, if not impossible, to map each title as a web citation. There is no direct way to select, reuse or quote content. Wikipedia and its sister projects, especially Wikidata, are impacted due to such technical restrictions. The number of online citations in Odia is already extremely low, as visible on Odia Wikipedia. The lack of content search prohibits Wikipedia editors from directly searching relevant content online for paraphrasing and citing. Additionally, files on Google can be subject to takedown and account and IP blocks. These challenges required re-archiving the already public texts in a stable, open and public archive.

Re-Archiving Process and Outcome

The Internet Archive (https://archive.org), a free and open digital archive with technical integrations with Wikimedia projects, was chosen to re-archive OB and OVA-archived texts. A file integration of Servants of Knowledge (SOK), a collaborator of Public.Resource.Org, helped re-archive the OB- and OVA-archive texts as scanned PDFs. In 2021, a small pilot was run to re-archive about 2,100 titles using the human- and machine-readable bilingual (English and Odia) metadata I created by expanding on any existing metadata. This pilot helped me create approximately 3,500 new Wikidata entries[1] about these books and their authors, enhanced with (Internet) Archive IDs for citation and authority control data such as Virtual International Authority File (VIAF). This Wikidata batch also includes 101 books and periodicals that were archived in a 2021 project collaborated on by Odia Wikimedians User Group, a Wikimedia affiliate, and Srujanika (odi, 2021). These Wikidata entries help Wikipedia editors identify notable authors missing from Wikipedia. For instance, a value of ”0” in the ”sites” column in a Wikidata query service for Odia-language authors (https://w.wiki/AM25) indicates the absence of a non-Wikidata Wikimedia project entry. During 2023 and 2024, I created metadata for the remaining 12,000+ books and periodicals, which SOK used for re-archiving. In total, 14,292 titles were re-archived, barring some duplicates yet to be identified and deleted by SOK. [2] From this, nearly 2300 books are in Public Domain in India that can potentially be digitised on Wikisource. By default, the Archive extracts text from scanned PDFs using Tesseract (Wajer, 2020), an open-source software, which is indexed on the web and allows any user to find the full text of a book through a mere search. Each book title or periodical issue has a unique Archive page, allowing one to cite in Wikimedia projects using Zotero (https://zotero.org) easily. The accuracy of the auto-generated text is yet to be computed, while its availability has addressed the primary purpose of re-archiving—making the books discoverable for the readers.

Identified Gaps

Auto-generated text has varied accuracy, depending on the scan quality and fonts used in typesetting. A handful of digital typefaces were used for the last time in 2019 for Tesseract training in Odia (Weil and Breidenbach, 2019), and Tesseract needs to be trained for letterpress typefaces to extract text from older books better. Wikidata entries on all the newly re-archived texts and the corresponding authors are yet to be created after matching existing entries and deduplicating them. Though re-archived books are already publicly available and mirrored online and are purely available in an archive, any potential copyright claim adversely impacting OB or OVA would also, in turn, impact the books' availability in the archive.

Discussion/Conclusions

Poor access to published knowledge and digital infrastructure can hinder universal access to knowledge in different language contexts, as in Odia. Such an issue also hampers projects like Wikipedia, which rely on volunteers for content creation and verification, resulting in lesser public and openly licensed encyclopaedic content. On the other hand, Scanned books have a better chance of being discoverable by users and Wikimedia contributors if the content text can be extracted and indexed on the web; the lack of this makes texts search-prohibitive. The availability of over 14,000 texts could help Odia readers who do not have access to physical libraries and archives and Wikimedia contributors to create and enhance Wikipedia entries and citations and digitise compatible openly licensed texts. Collaboration between digital libraries/ archives and Wikimedians is essential to identify shared gaps and action. OCR is not uniform across languages, and more collaboration between technical contributors and organisations is required to improve open-source OCR engines such as Tesseract. That would significantly improve the image-to-text extraction quality. Wikidata is a useful bridge between the Internet Archive and other Wikimedia projects. It helps capture metadata about notable writers and books, and Wikidata queries help identify notable writers without Wikipedia entries.

References

[Mishra2017] Sweta Mishra. 2017. Odia Bibhaba: A treasure trove of Odia literature. *Sambad English*, November.

[odi2021] 2021. Odia Wikimedians User Group/Book digitization 2021 - Meta, July.

[Panigrahi2022] Subhashish Panigrahi. 2022. The Volunteer Archivists, May. Place: Bengaluru OCLC: 1333435715 QID: Q113199297.

[Siksha ’O’ Anusandhan2021] Siksha ’O’ Anusandhan. 2021. About PPRACHIN, October.

[Srujanika2024] Srujanika. 2024. Odia Bibhaba | Treasure Trove of Things Odia, April.

[Wajer2020] Merlijn Wajer. 2020. OCR at the Internet Archive with Tesseract and hOCR — Internet Archive Developer Portal, January.

[Weil and Breidenbach2019] Stefan Weil and Jeff Breidenbach. 2019. langdata_lstm/ori/okfonts.txt at main · tesseract-ocr/langdata_lstm, October.

[1] An Odia-language author list created using the ListeriaBot.
[2] See re-archived texts' list here.