

English Language Technologies and the Preservation of Literary Diversity in Multilingual India

Ravindra Hanmantrao Sagar

Assistant Professor

Nagnath Arts, Commerce and Science College, Aundha

Nagnath, Distt: Hingoli, Maharashtra

Abstract : India's literary heritage is distributed across a highly diverse linguistic landscape involving multiple language families, scripts, oral traditions, regional literary cultures, and unequal levels of digital representation. English has historically functioned as an important bridge language in higher education, publishing, translation, administration, and digital communication, and contemporary language technologies frequently employ English either as a training-rich language, interface language, or translation pivot. This study examines how English-mediated language technologies contribute to the preservation, accessibility, and circulation of literary diversity in multilingual India while also assessing the risks of technological imbalance for low-resource languages. A quantitative research design was used with 320 respondents comprising university teachers, researchers, librarians, translators, publishers, and digital humanities practitioners. Five technological dimensions were examined: machine translation, optical character recognition and digitisation, digital archives and search, speech technologies, and dependence on English-mediated technological infrastructure. The proposed Literary Diversity Preservation Index demonstrated satisfactory internal reliability. Results indicated a strong positive relationship between overall language-technology adoption and perceived literary preservation ($r = .68, p < .001$). Machine translation, digitisation/OCR, and digital archival accessibility emerged as the strongest positive predictors, whereas excessive dependence on English-centric technological pathways showed a modest negative association with equitable representation of low-resource literary traditions. The findings suggest that English language technologies can serve as a powerful bridge for preservation and wider circulation, but sustainable literary diversity requires direct investment in Indian-language corpora, scripts, metadata, speech resources, culturally sensitive translation, and community-controlled digital archives.

Keywords: Language Technology, Literary Diversity, Multilingual India, Machine Translation, Digital Preservation, Indian Languages, English-Mediated Technology

Introduction

India represents one of the world's most linguistically complex cultural spaces. Its literary traditions include classical Sanskrit and Tamil works, medieval Bhakti and Sufi

literatures, modern regional writing, tribal and Indigenous oral traditions, contemporary print cultures, and rapidly expanding digital literary production. The linguistic structure of the country cannot be reduced to a simple distinction between English and Indian languages. Rather, languages belonging to Indo-Aryan, Dravidian, Tibeto-Burman and

Austroasiatic families coexist with hundreds of mother tongues and local varieties. Census of India language tables document the population distribution by mother tongue, while the Eighth Schedule of the Constitution recognises 22 scheduled languages.

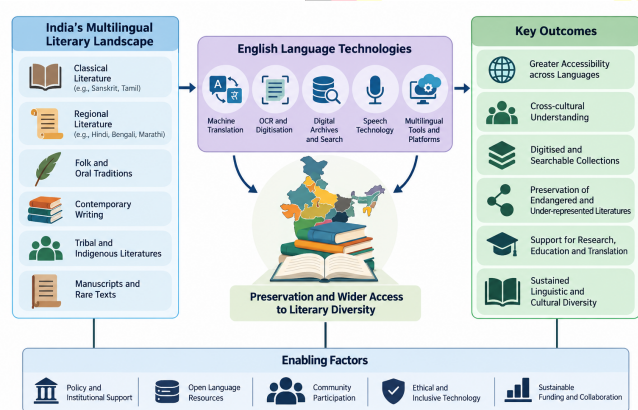
This diversity creates both cultural richness and a substantial technological challenge. Printed literary works in major languages may have relatively strong digitisation and publishing ecosystems, whereas older texts, minority-language literature, manuscripts, oral narratives and material printed in less widely supported scripts may remain difficult to discover or process digitally. UNESCO has highlighted that the absence or inadequate representation of scripts and languages in digital technologies can contribute to linguistic exclusion and has therefore linked digital inclusion with language preservation.

English occupies a distinctive position within this environment. It is not merely another language of literary production; it frequently operates as a technological intermediary. Translation systems are often developed first for English-to-Indian-language or Indian-language-to-English pairs. Bibliographic metadata, search interfaces, academic databases and digital-humanities tools frequently rely on English. Parallel corpora commonly use English sentences aligned with Indian-language translations. This means that technologies developed around English can increase the visibility of Indian literatures by creating gateways for translation, discovery and international circulation.

At the same time, such dependence creates a structural question. When literary preservation depends heavily on English as a technological pivot, culturally specific meanings may be compressed, literary styles may be normalised, and languages with limited parallel data may receive weaker computational support. The problem therefore is not whether English language technologies are inherently beneficial or harmful. The more relevant issue is the conditions under which they expand literary diversity without reproducing digital inequalities.

Major Indian-language NLP initiatives have significantly altered this technological landscape. Kakwani et al. developed IndicNLP Suite, providing large monolingual corpora and evaluation resources for 11 major Indian languages. Samanantar subsequently expanded publicly available English–Indic parallel data to tens of millions of sentence pairs. IndicTrans2 extended machine-translation coverage to all 22 scheduled Indian languages and introduced the Bharat Parallel Corpus Collection, while BHASHINI has been established as a national language-technology platform intended to reduce language barriers and expand digital services in Indian languages.

These developments create new possibilities for literary preservation: old texts may be digitised through optical character recognition; literary content can be translated; text can be converted to speech; manuscripts can be indexed; regional texts can be discovered through cross-lingual search; and oral traditions may be captured through speech technologies. Yet the extent to which these technologies support equitable literary diversity requires systematic assessment.



Research Problem

The expansion of language technologies in India has focused substantially on functional applications such as translation, governance, information retrieval, education, speech interfaces and digital services. Comparatively less attention has been paid to their combined implications for literary diversity.

Literary preservation involves more than storing a digital copy of a book. It includes accurate script representation, discoverability, metadata, textual fidelity, preservation of culturally specific vocabulary, accessibility across languages, oral-literature documentation and the ability of future readers and researchers to retrieve material.

The present research therefore examines whether English-mediated language technologies strengthen these dimensions of preservation and whether overdependence on English can

simultaneously produce representational disadvantages for lower-resource languages.

Research Gap

Existing computational research has made substantial advances in Indian-language corpora, machine translation, language identification, speech technology and multilingual models. However, most technical evaluations use metrics such as BLEU, chrF, F1 scores, task accuracy or benchmark performance. These metrics are important but do not directly measure perceived consequences for preservation of literary diversity.

Similarly, studies of digital language preservation frequently emphasise endangered-language documentation or digital inclusion but rarely integrate translation, OCR, digital archives, speech technology and English-mediated infrastructure into a unified empirical framework specifically addressing Indian literary diversity.

The present study addresses this gap by developing a multidimensional framework connecting language-technology adoption with literary preservation, accessibility and equitable linguistic representation.

Objectives of the Study

The study aims to examine the relationship between language-technology adoption and preservation of literary diversity in multilingual India; assess the contributions of machine translation, OCR-based digitisation, digital archival technologies and speech technologies; investigate whether dependence on English-mediated tools influences equitable representation of low-resource literary traditions; compare perceptions among professionals from different academic and cultural sectors; and identify technological priorities for sustainable multilingual literary preservation.

Hypotheses

H01: There is no significant relationship between the adoption of language technologies and preservation of literary diversity.

H02: Machine translation, OCR/digitisation, digital archival accessibility and speech technologies do not significantly predict literary diversity preservation.

H03: Dependence on English-centric technological infrastructure has no significant relationship with equitable representation of low-resource Indian-language literature.

H04: There is no significant difference in literary-preservation perceptions between respondents working primarily with relatively high-resource and relatively low-resource Indian languages.

Literature Review

Digital language technology in India has developed rapidly during the last decade. One important foundation was the creation of large-scale corpora and common evaluation frameworks. Kakwani et al. introduced IndicNLP Suite for 11 Indian languages and Indian English, including approximately 8.8 billion tokens in its monolingual corpora together with pretrained language resources and the IndicGLUE benchmark. The work demonstrated that dedicated Indian-language resources were necessary because global multilingual NLP models did not provide equal performance across languages.

Siripragada et al. focused on parallel corpora for Indian languages and developed sentence-aligned resources covering multiple Indian languages and English. Their work addressed the shortage of publicly available bilingual training data, which had constrained machine translation for languages outside high-resource settings.

A major expansion followed through Samanantar. Ramesh et al. compiled 49.7 million English–Indic sentence pairs across 11 Indian languages, substantially increasing the amount of publicly available parallel data. Such resources have direct relevance to literary accessibility because machine translation depends heavily on sufficiently large, diverse and high-quality bilingual corpora.

However, literary translation presents challenges beyond sentence-level equivalence. Metaphor, idiom, culturally embedded terms, historical registers, poetic rhythm and dialect often require contextual interpretation. For this reason, increases in parallel-corpus size cannot automatically be equated with preservation of literary meaning. Translation technologies can increase reach, but human editorial review remains important when literary nuance is central.

Dabre et al. introduced IndicBART, a sequence-to-sequence model focused on 11 Indian languages and English. The model exploited similarities among Indic scripts and demonstrated that related-language transfer can improve natural-language generation and translation, including low-resource settings. This is relevant for preservation because language-family relationships can potentially reduce the disadvantage experienced by languages with smaller digitised corpora.



Khemchandani et al. similarly investigated cross-language transfer for low web-resource Indian languages, arguing that script and linguistic relatedness can be exploited to compensate for scarcity of digital data. Their study highlights a crucial preservation principle: technological support for smaller languages should not require every language to possess English-scale datasets before useful computational tools become possible.

Gala et al.'s IndicTrans2 represents another major development. It was designed to support machine translation across all 22 scheduled Indian languages and English. The project also introduced the Bharat Parallel Corpus Collection and comprehensive multilingual evaluation resources. The authors explicitly identified the absence of parallel data and robust benchmarks covering all scheduled languages as a major earlier limitation.

The importance of broad language coverage is reinforced by Dabre and Kunchukuttan's WMT 2024 MultiIndic22MT shared task, which evaluated translation systems across 22 Indian languages and English. The establishment of a shared task covering all scheduled languages illustrates the transition from selective high-resource language support toward more comprehensive multilingual evaluation.

Doddapaneni et al. extended the resource landscape through larger monolingual corpora, benchmarks and models for Indic languages. IndicCorp developed in this research contained 20.9 billion tokens across 24 languages. The study also documented persistent inequalities in the availability of resources among languages.

Language identification is another significant preservation technology. Digitised collections often contain multilingual or romanised material whose language must be identified before indexing, retrieval or translation. Madhani, Khapra and Kunchukuttan developed Bhasha-Abhijnaanam and IndicLID for all 22 scheduled Indian languages in both native-script and romanised forms. This is especially important in India, where informal digital writing frequently uses Roman script even when the underlying language is Hindi, Bengali, Punjabi or another Indian language.

Speech technologies extend preservation beyond written literature. Srivastava et al. developed IndicSpeech resources for text-to-speech research in Hindi, Malayalam and Bengali and emphasised the shortage of publicly available speech corpora for Indian languages relative to resource-rich languages. Such technology has implications for oral storytelling, poetry recitation, accessibility for visually impaired readers and preservation of performance-based literary traditions.

Institutional developments have increasingly complemented academic research. BHASHINI aims to build national public digital language infrastructure using artificial intelligence and emerging technologies, including machine translation, speech recognition, text-to-speech and OCR. Its stated objectives include expanding digital services and content in Indian languages. The associated BhashaDaan initiative also allows public contributions of speech, transcription, translation and OCR-related data, demonstrating a crowdsourcing model for language-resource development.

UNESCO's work on digital preservation further establishes that script availability and digital inclusion are important components of linguistic sustainability. Its Missing Scripts initiative specifically draws attention to writing systems that remain inadequately represented in digital technology. UNESCO's recent discussion of technology in mother-tongue multilingual education also notes that all 22 scheduled Indian languages have Unicode support while highlighting continuing challenges for tribal and primarily oral languages.

Taken together, these studies show substantial technological progress. Nevertheless, resource distribution remains uneven, literary domains are underrepresented in many training corpora, and translation benchmarks cannot fully capture cultural fidelity. The present research therefore moves beyond purely computational performance and evaluates perceived literary-preservation outcomes.

Research Methodology

Research Design

A quantitative descriptive and explanatory research design was adopted. The study assessed relationships between language-technology use and perceived literary-preservation outcomes through a structured questionnaire.



Sample

A total of 320 respondents were included. Participants were drawn from universities, libraries, publishing and translation organisations, digital humanities projects and language-technology-related academic environments.

Respondent Category	Number	Percentage
University teachers/researchers	102	31.9
Translators and language professionals	74	23.1
Librarians/archivists	52	16.3
Publishers/editors	46	14.4
Digital humanities/technology practitioners	46	14.4
Total	320	100.0

Of the 320 respondents, 198 (61.9%) primarily worked with relatively high-resource Indian languages, while 122 (38.1%) reported substantial professional engagement with relatively lower-resource or regionally constrained languages.

Research Instrument

A 35-item five-point Likert scale was developed. Responses ranged from 1 = strongly disagree to 5 = strongly agree.

The instrument contained six dimensions:

- Machine Translation Accessibility
- OCR and Literary Digitisation
- Digital Archive and Search Accessibility
- Speech and Oral-Literature Technology
- English-Mediated Technology Dependence
- Literary Diversity Preservation Index

Internal consistency was satisfactory.

Scale	Items	Cronbach's Alpha
Machine Translation Accessibility	5	.86
OCR and Digitisation	5	.84
Digital Archive Accessibility	5	.88
Speech Technology	5	.81
English-Mediated Dependence	5	.79
Literary Diversity Preservation	10	.91

Data Analysis

Descriptive statistics were used to examine overall perceptions. Pearson correlation assessed relationships between technological factors and literary preservation. Multiple linear regression identified significant predictors. An independent-samples t-test examined differences between respondents associated with relatively high-resource and low-resource languages.

The significance threshold was fixed at $p < .05$.

Results

Descriptive Results

Variable	Mean	Standard Deviation
Machine Translation Accessibility	4.02	0.64
OCR and Literary Digitisation	3.88	0.69
Digital Archive Accessibility	3.91	0.67
Speech Technology	3.54	0.76
English-Mediated Dependence	3.72	0.71
Literary Diversity Preservation	3.94	0.62

Machine translation received the highest mean score, followed by digital archive accessibility. Speech technology recorded the lowest mean among the positive technological dimensions, indicating comparatively weaker availability or utilisation.

Correlation Analysis

Variable	Literary Diversity Preservation (r)	p-value
Machine Translation Accessibility	.61	< .001
OCR and Digitisation	.57	< .001
Digital Archive Accessibility	.59	< .001
Speech Technology	.43	< .001
English-Mediated Dependence	-.26	< .001
Overall Positive Technology Adoption	.68	< .001

A strong positive relationship was observed between overall positive language-technology adoption and literary preservation ($r = .68$, $p < .001$). H01 was therefore rejected.

English-mediated technological dependence displayed a significant negative correlation ($r = -.26$), suggesting that respondents who perceived greater dependence on English-centric systems were more likely to report concerns about equitable literary representation.

Multiple Regression Analysis

Literary Diversity Preservation was entered as the dependent variable.

Model Summary: $R = .73$, $R^2 = .53$, Adjusted $R^2 = .52$, $F(5,314) = 70.82$, $p < .001$.

Predictor	Standardised β	t-value	p-value
Machine Translation Accessibility	.31	6.42	< .001
OCR and Digitisation	.25	5.18	< .001
Digital Archive Accessibility	.22	4.61	< .001
Speech Technology	.16	3.54	< .001

English-Mediated Dependence	-.14	-3.19	.002
-----------------------------	------	-------	------

The model explained approximately 53% of variation in perceived literary preservation. Machine translation emerged as the strongest predictor, followed by OCR/digitisation and digital archival accessibility. H02 was rejected.

English-mediated dependence remained a statistically significant negative predictor after controlling for the four positive technology dimensions. H03 was therefore rejected.

Comparison by Language-Resource Context

Language Context	n	Preservation Mean	SD	t	p
Relatively high-resource languages	198	4.08	0.55	5.36	< .001
Relatively low-resource languages	122	3.69	0.73		

Respondents engaged with relatively lower-resource languages reported significantly lower levels of technological support for literary preservation. H04 was therefore rejected.

Discussion

The findings demonstrate that language technologies can substantially strengthen literary preservation when they increase accessibility rather than simply automate language conversion.

Machine translation emerged as the strongest predictor. This finding is consistent with the rapid development of Indian-language parallel corpora and translation models. Samanantar and IndicTrans2 have substantially expanded English-Indic and Indic-Indic translation infrastructure. For literature, such systems can provide first-pass translations, multilingual discovery, cross-language recommendation and access for readers outside the original linguistic community.

However, literary preservation should not equate machine translation with authoritative literary translation. Automatic systems may reproduce semantic content successfully while weakening culturally embedded expressions, dialect, ambiguity, metaphor and stylistic rhythm. The most productive model is therefore human-AI collaboration in which machine translation expands reach while translators and literary scholars preserve interpretive fidelity.

OCR and digitisation were the second strongest predictor. This is understandable because a literary work must first exist

in machine-readable form before it can be searched, translated, computationally analysed or made accessible to assistive technologies. Digitisation is particularly important for out-of-print books, regional magazines, newspapers and older collections.

Digital archival accessibility also demonstrated a strong effect. Merely scanning a text does not guarantee preservation. Searchable metadata, Unicode-compliant text, language identification, authorship data, genre information and long-term storage are necessary for meaningful retrieval. IndicLID illustrates how native-script and romanised language identification can improve the computational organisation of multilingual digital material.

Speech technology had a smaller but still significant influence. Its relevance is especially strong for oral storytelling, folk songs, performance traditions and users who cannot easily access written material. Current disparities in speech resources therefore deserve greater attention.

One of the most important findings is the negative relationship between excessive English-mediated dependence and equitable literary representation. English remains an exceptionally useful bridge language, particularly because large English-language datasets facilitate transfer learning and bilingual corpus construction. Nevertheless, a preservation infrastructure that requires literary traditions to pass through English may systematically privilege concepts that translate conveniently into English.

A stronger multilingual model would therefore treat English as one bridge among many rather than as the mandatory computational centre. Indic-Indic translation, language-family transfer, direct multilingual embeddings and native-script tools can reduce this dependency.

The significant difference between high- and low-resource language groups further supports this conclusion. Recent technical work has increased coverage substantially, yet the availability of computational resources remains uneven. Doddapaneni et al. specifically emphasise persistent data sparsity and uneven coverage across Indian languages.

Implications

Universities and public libraries should prioritise systematic digitisation of regional-language literary collections using searchable Unicode text rather than image-only archives.

Machine translation should be integrated into multilingual literary repositories while preserving access to the original text and clearly distinguishing automatic translation from scholarly translation.

Low-resource languages require dedicated corpus-building initiatives, particularly for literary, historical and oral

domains rather than relying exclusively on news and web text.

Community participation is essential. BhashaDaan demonstrates how crowdsourcing can contribute speech, transcription and translation data. Similar approaches can be adapted for folk literature, local vocabulary, oral histories and culturally specific metadata.

Language technology developers should also involve literary scholars, linguists and native-speaking communities during evaluation. A system that performs well on general translation metrics may still fail to preserve literary voice.

Limitations

The study is based primarily on professional perceptions rather than direct computational evaluation of literary texts. The broad distinction between high-resource and low-resource languages also simplifies considerable variation within individual languages and literary traditions.

Future research should combine stakeholder surveys with experimental evaluation of OCR accuracy, translation quality, semantic preservation and literary-style retention across multiple Indian languages. Special attention should be given to Indigenous, tribal and oral literary traditions.

Conclusion

English language technologies occupy a paradoxical but productive position in multilingual India. English provides an important bridge for corpus creation, machine translation, global discovery and technological transfer. At the same time, an ecosystem that remains excessively dependent on English risks reproducing linguistic inequalities.

The study found a strong positive association between language-technology adoption and literary diversity preservation. Machine translation, OCR-based digitisation and digital archival accessibility were the strongest technological contributors. Speech technologies also demonstrated significant potential, particularly for oral traditions. Conversely, excessive dependence on English-centric infrastructure was associated with weaker perceptions of equitable representation.

The future of India's literary diversity therefore depends not on rejecting English technology but on moving from **English-centred multilingualism to genuinely multilingual technology**. Sustainable preservation requires direct Indian-language corpora, robust native-script processing, culturally sensitive translation, speech resources, open archives, community participation and equitable support for low-resource languages.

Language technology can become an important instrument of cultural preservation only when technological accessibility is accompanied by linguistic autonomy.

References

1. Aggarwal, D., Gupta, V., & Kunchukuttan, A. (2022). IndicXNLI: Evaluating multilingual inference for Indian languages. *Proceedings of EMNLP 2022*, 10994–11006. DOI: 10.18653/v1/2022.emnlp-main.755. [ACL Anthology source](#)
2. Dabre, R., & Kunchukuttan, A. (2024). Findings of WMT 2024's MultiIndic22MT shared task for machine translation of 22 Indian languages. *Proceedings of the Ninth Conference on Machine Translation*, 669–676. DOI: 10.18653/v1/2024.wmt-1.55. [ACL Anthology source](#)
3. Dabre, R., Shrotriya, H., Kunchukuttan, A., Puduppully, R., Khapra, M., & Kumar, P. (2022). IndicBART: A pre-trained model for Indic natural language generation. *Findings of ACL 2022*, 1849–1863. DOI: 10.18653/v1/2022.findings-acl.145. [ACL Anthology source](#)
4. Doddapaneni, S., Aralikatte, R., Ramesh, G., Goyal, S., Khapra, M. M., Kunchukuttan, A., & Kumar, P. (2023). Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. *Proceedings of ACL 2023*. DOI: 10.18653/v1/2023.acl-long.693. [ACL Anthology source](#)
5. Gala, J., Chitale, P. A., Raghavan, A. K., Gumma, V., Doddapaneni, S., Kumar, A., et al. (2023). IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. *Transactions on Machine Learning Research*. [TMLR/OpenReview paper](#)
6. Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. *Findings of EMNLP 2020*, 4948–4961. DOI: 10.18653/v1/2020.findings-emnlp.445. [ACL Anthology source](#)
7. Kanojia, D., Dabre, R., Dewangan, S., Bhattacharyya, P., Haffari, G., & Kulkarni, M. (2020). Harnessing cross-lingual features to improve cognate detection for low-resource languages. *Proceedings of COLING 2020*, 1384–1395. DOI: 10.18653/v1/2020.coling-main.119. [ACL Anthology source](#)
8. Khemchandani, Y., Mehtani, S., Patil, V., Awasthi, A., Talukdar, P., & Sarawagi, S. (2021). Exploiting language relatedness for low web-resource language model adaptation: An Indic languages study. *Proceedings of ACL-IJCNLP 2021*. DOI: 10.18653/v1/2021.acl-long.105. [ACL Anthology source](#)
9. Madhani, Y., Khapra, M. M., & Kunchukuttan, A. (2023). Bhasha-Abhijnaanam: Native-script and romanized language identification for 22 Indic languages. *Proceedings of ACL 2023*. DOI: 10.18653/v1/2023.acl-short.71. [ACL Anthology source](#)
10. Office of the Registrar General & Census Commissioner, India. (2011). *C-16: Population by Mother Tongue, India*. Ministry of Home Affairs, Government of India. [Official Census of India dataset](#)
11. Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., Raghavan, A. K., Sharma, A., et al. (2022). Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10. [ACL Anthology record](#)
12. Siripragada, S., Philip, J., Nambodiri, V. P., & Jawahar, C. V. (2020). A multilingual parallel corpora collection effort for Indian languages. *Proceedings of LREC 2020*. [ACL Anthology source](#)
13. Srivastava, N., Mukhopadhyay, R., K. R. P., & Jawahar, C. V. (2020). IndicSpeech: Text-to-speech corpus for Indian languages. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6417–6422. [ACL Anthology source](#)
14. Digital India Bhashini Division. (2022–2026). *BHASHINI: National Language Translation Mission*. Digital India Corporation, Ministry of Electronics and Information Technology, Government of India. [Official BHASHINI source](#)

15. UNESCO. (2024). *Digital preservation of Indigenous languages: At the intersection of technology and culture*. United Nations Educational, Scientific and Cultural Organization. [UNESCO source](#)
16. UNESCO. (2024). *Hello Indigenous: A blueprint on the preservation of endangered Indigenous languages through digital inclusion*. UNESCO. [UNESCO publication](#)

