# Digitalizing the Spoken Word : AI and Big Data Frameworks for Preserving Oral Traditions in Rural Development.


## Abstract

Oral traditions constitute a foundational layer of cultural identity, indigenous knowledge, and community governance in rural societies worldwide, yet they remain acutely vulnerable to erosion through generational discontinuity, migration, and the displacement effects of globalization. This study presents an integrated framework combining artificial intelligence and big data technologies to systematically capture, classify, and preserve oral heritage as a driver of sustainable rural development. Drawing on a qualitative method, the research employs automatic speech recognition (ASR), natural language processing (NLP), and deep learning algorithms to transcribe, translate, and semantically annotate spoken narratives, folk knowledge, and traditional ecological practices in low-resource languages. A cloud-based big data architecture was designed to manage heterogeneous multimedia corpora at scale, ensuring interoperability with global digital heritage repositories. Findings demonstrate that AI-assisted digitalization achieves transcription with higher accuracy while substantially reducing archival processing time compared to manual methodologies. Critically, the framework incorporates community-governed data sovereignty protocols, ensuring that digitalized content remains ethically anchored to its source communities and supports locally led rural development agendas, including agricultural knowledge transfer, indigenous health practices, and customary resource management. The study argues that responsible AI deployment in oral heritage preservation is not merely a technological endeavor but a socio-political act of epistemic justice, recentering marginalized knowledge systems within broader development discourses.



**Keywords:** oral traditions preservation; artificial intelligence; big data; rural development; indigenous knowledge systems; low-resource languages; data sovereignty


## 1. Introduction

Oral traditions (proverbs, epics, genealogies, praise songs, ritual chants, agricultural calendars, medicinal recitations and narratives of origin) form the principal knowledge infrastructure of many rural societies, particularly where literacy in official languages is uneven and writing has historically played a secondary role in transmitting knowledge (Finnegan, 2012; Hampâté Bâ, 1981; Vansina, 1985). They carry more than aesthetic value. They encode land-tenure arrangements, genealogical claims, rules of dispute settlement, ecological observation and therapeutic practice, and in doing so they sustain the cultural identity and governance capacity of the communities that hold them. The 2003 UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage recognises oral traditions and expressions, including language as their vehicle, as a core domain of living heritage

UNDER PEER REVIEW IN IJAR

(UNESCO, 2003). The oft-cited remark attributed to Amadou Hampâté Bâ, that the death of an elder in Africa is comparable to the burning of a library, captures the fragility of this heritage: when a custodian dies without having transmitted a repertoire, that repertoire may disappear with him or her.

This fragility is intensifying. Rural-to-urban and transnational migration separates younger generations from the settings in which oral repertoires are performed and learned. Formal schooling, administration and mass media operate mostly in a small number of dominant languages. Climatic stress and insecurity displace entire communities and interrupt seasonal ceremonies and apprenticeships. Linguists have warned for three decades that the erosion of minority languages carries away the knowledge systems expressed in them (Krauss, 1992; Moseley, 2010). Preservation is therefore not only a cultural concern. Where indigenous knowledge about crops, pasture, water, medicinal plants and conflict mediation is lost, rural communities lose resources that development practice increasingly acknowledges as valuable (Sillitoe, 1998; Warren et al., 1995; World Bank, 1998).

Documentation has long been constrained by a labour bottleneck: transcribing, translating and annotating a recording takes many times its length and requires scarce expertise (Himmelmann, 1998), so archives hold large volumes of audio that remain unsearchable and, in practice, unusable by the communities that produced it. Advances in artificial intelligence (AI) promise to relax this constraint. Self-supervised speech representations can be learned from unlabelled audio and adapted with little transcribed data (Baevski et al., 2020; Babu et al., 2022); multilingual systems now cover more than a thousand languages for recognition (Pratap et al., 2024); weakly supervised models transcribe close to a hundred (Radford et al., 2023); machine translation has been extended to two hundred (NLLB Team, 2024); and Masakhane has shown that African-language technology can be built participatively (Nekoto et al., 2020).

Yet three gaps remain. First, technological performance is unevenly distributed : languages with little digital text or labelled speech continue to be served poorly (Blasi et al., 2022; Joshi et al., 2020). Second, much technical work treats heritage as a data-processing problem and neglects the infrastructure needed to store, curate and interlink heterogeneous multimedia corpora at scale and to keep them interoperable with global repositories. Third, and most consequentially, the governance question (who owns, controls, accesses and benefits from digitalized oral heritage) is often treated as an afterthought, even though critical

UNDER PEER REVIEW IN IJAR

scholarship has warned that extractive data practices can reproduce colonial relations in digital form (Birhane, 2020; Couldry & Mejias, 2019; Mohamed et al., 2020).

This article addresses these gaps by proposing and discussing an integrated framework that combines AI and big data technologies with community-governed data sovereignty protocols. Four research questions guide the study. (RQ1) How can automatic speech recognition (ASR), natural language processing (NLP) and deep learning be configured to transcribe, translate and semantically annotate oral heritage in low-resource languages? (RQ2) What big data architecture can manage heterogeneous multimedia corpora at scale while remaining interoperable with global digital heritage repositories? (RQ3) What governance protocols keep digitalized content ethically anchored to its source communities? (RQ4) How can the resulting digital resources support locally led rural development agendas, notably agricultural knowledge transfer, indigenous health practices and customary resource management?

## 2. Conceptual and Literature Background

### 2.1 Oral traditions as Living Knowledge Systems

Ong (1982) characterised primary oral cultures as organised around memory, formulaic expression, repetition and the performance situation rather than around fixed text. Finnegan (2012) stressed that African oral literature cannot be reduced to its words: music, gesture, audience response, setting and the identity of the performer are constitutive of meaning. Vansina (1985) approached oral tradition as a historical source transmitted along chains of tellers, with each link shaping the message, while Hampâté Bâ (1981) emphasised that the living tradition is inseparable from an ethics of transmission, with obligations placed on the transmitter, restrictions on who may speak what, and an integration of the sacred, the practical and the social.

Three implications follow for digitalization. Transcripts alone cannot stand for a performance, so capture must be multimodal. Oral traditions are variable by nature, so a single ‘authoritative’ text risks freezing a fluid practice. And some knowledge is restricted by age, gender, initiation or ritual status, so access cannot be uniform.

### 2.2 Indigenous Knowledge and Rural Development

Since the 1990s, development institutions have moved from treating local knowledge as an obstacle to modernisation to treating it as a resource for participatory and sustainable

UNDER PEER REVIEW IN IJAR

development (Chambers, 1994; World Bank, 1998). Warren et al. (1995) documented the role of indigenous knowledge systems in agriculture, resource management and health across several regions. Berkes (2012) conceptualised traditional ecological knowledge as a cumulative body of knowledge, practice and belief transmitted culturally across generations. Critics caution against romanticising or reifying such knowledge and against drawing a sharp line between ‘indigenous’ and ‘scientific’ knowledge (Agrawal, 1995; Sillitoe, 1998). These cautions matter for system design : the framework presented here treats oral knowledge as dynamic, contested and context-dependent, and positions digital resources as support for community deliberation rather than as replacement for it.

### 2.3 Language Documentation and its Labour Bottleneck

Documentary linguistics reoriented the field towards lasting, multipurpose records of linguistic practice, comprising recordings and their annotation, translation and metadata (Himmelmann, 1998). Bird and Simons (2003) proposed criteria for the portability of such records across time, software and community, and repositories such as the Endangered Languages Archive (ELAR) and PARADISEC embody them. Even so, few recordings are fully transcribed and translated, because manual work is slow and depends on a handful of speakers who are also teachers, farmers and elders with other obligations. AI-assisted workflows address this problem, but it is also where the risks of automation and extraction are greatest.

### 2.4 AI for Low-Resource Languages

Modern ASR has moved from hybrid pipelines to end-to-end neural models based on attention architectures (Vaswani et al., 2017). The decisive development for low-resource settings is self-supervised pre-training: wav2vec 2.0 learns representations from raw audio and reaches useful accuracy after fine-tuning on small labelled sets (Baevski et al., 2020). Cross-lingual extensions (Babu et al., 2022) and the Massively Multilingual Speech project (Pratap et al., 2024) widened language coverage, while Whisper demonstrated the robustness gained from large-scale weak supervision (Radford et al., 2023). Parameter-efficient adaptation methods such as low-rank adaptation (Hu et al., 2022) make it possible to specialise large models on modest hardware, which is important for institutions in the Global South.

In text processing, multilingual encoders (Conneau et al., 2020), African-language models trained on small corpora (Ogueji et al., 2021) and translation systems covering many African

UNDER PEER REVIEW IN IJAR

languages (NLLB Team, 2024) provide building blocks for semantic annotation, while multilingual embeddings enable semantic search across transcripts and translations. Nonetheless, evaluations repeatedly show systematic inequalities : Joshi et al. (2020) classified most of the world's languages as effectively invisible to NLP, Blasi et al. (2022) quantified the gap in performance across languages, and Koenecke et al. (2020) showed that commercial ASR performs worse for some speaker groups. Generative recognisers can also 'hallucinate' text that was never spoken, with potential for harm in sensitive contexts (Koenecke et al., 2024). These findings justify the framework's insistence on human verification and transparent documentation, in the spirit of data statements (Bender & Friedman, 2018), datasheets for datasets (Gebru et al., 2021) and model cards (Mitchell et al., 2019).

### 2.5 Big Data Infrastructures for Heritage

Digitalized oral heritage displays the classic characteristics of big data : volume, variety (audio, video, transcripts, translations, images, field notes) and velocity of ingestion from distributed contributors. Distributed processing such as MapReduce (Dean & Ghemawat, 2008) and Spark (Zaharia et al., 2016) allows batch transcription, embedding and indexing at scale. For long-term value, data should be findable, accessible, interoperable and reusable (FAIR; Wilkinson et al., 2016) and follow community metadata standards such as those of the Open Language Archives Community (Simons & Bird, 2003). FAIR alone is insufficient for heritage that belongs to communities, and must be complemented by the CARE principles of collective benefit, authority to control, responsibility and ethics (Carroll et al., 2020).

### 2.6 Data Sovereignty and Epistemic Justice

Indigenous data sovereignty asserts the right of communities to govern the collection, ownership and application of data about their people, lands and cultures (Kukutai & Taylor, 2016; Rainie et al., 2019). The First Nations principles of ownership, control, access and possession (First Nations Information Governance Centre, 2014), the CARE principles (Carroll et al., 2020), and Article 31 of the United Nations Declaration on the Rights of Indigenous Peoples (United Nations General Assembly, 2007) provide normative reference points, while labelling initiatives such as Traditional Knowledge Labels illustrate how provenance and cultural protocols can be encoded in metadata (Anderson & Christen, 2013). On the African continent, instruments such as the Malabo Convention on cyber security and personal data protection (African Union, 2014) and national statutes such as Ghana's Data

UNDER PEER REVIEW IN IJAP

Protection Act (Republic of Ghana, 2012) frame the legal environment, although none specifically addresses collective rights over oral heritage.

Behind these instruments lies a broader argument about knowledge itself. Fricker (2007) defined epistemic injustice as a wrong done to someone in their capacity as a knower ; de Sousa Santos (2014) described the destruction of non-Western knowledge systems as epistemicide ; Smith (2012) criticised research practices that treat indigenous communities as objects of study. In African letters, Ngũgĩ wa Thiong'o (1986) argued that language functions as a collective memory bank carrying a people's historical experience, so that the displacement of African languages is a displacement of memory and self-definition. In this respect, Bird (2020) extended this critique to speech and language technology, calling for approaches that begin from the priorities of language communities. It is within this tradition that the present article frames AI-assisted digitalization as a socio-political act: depending on who controls it, it can either deepen epistemic marginalisation or help to recentre marginalised knowledge in development discourse.

## 3. Methodology

### 3.1 Research Design

The study adopts a qualitative, design-oriented approach appropriate to the construction and appraisal of a socio-technical framework (Lincoln & Guba, 1985). The design had four iterative components. First, a structured conceptual synthesis was conducted of scholarship on oral tradition, indigenous knowledge and development, language documentation, speech and language technology for low-resource languages, big data infrastructure, and indigenous data governance. Second, the technical and governance layers of the framework were specified from this synthesis. Third, the framework was appraised against representative genres of rural oral heritage (narratives and histories, proverbs and maxims, praise and ritual songs, and procedural knowledge on farming, healing and resource use) using criteria derived from the literature. Fourth, the resulting configurations were analysed thematically to identify enabling conditions, tensions and risks (Braun & Clarke, 2006).

### 3.2 Setting

The framework targets low-resource languages, those with little labelled speech or digital text, as is the case for most languages spoken in rural West Africa. The Sahel and the Gulf of Guinea host dense oral traditions, including epic and praise-singing, drum and instrumental

UNDER PEER REVIEW IN IJAR

texts, ritual chant, proverb performance, and agro-pastoral and ethnomedical lore, yet these languages have limited representation in speech corpora. The analysis is designed to be transferable to other rural contexts with similar conditions.

### 3.3 Evaluation Protocol and Indicators

To make the appraisal systematic, five families of indicators were specified (Table 1). Transcription quality is expressed through word error rate, WER = (S + D + I) / N, where S, D and I denote Substituted, Deleted and Inserted words and N the number of words in the reference transcript, complemented by character error rate, which is more informative for agglutinative or orthographically unstable languages. Efficiency is expressed as the ratio between human working time and recording duration, compared across manual and AI-assisted workflows. Annotation reliability is assessed through inter-annotator agreement using Cohen's (1960) kappa. Because numerical accuracy does not capture cultural adequacy, community validation (judgements by custodians and speakers of whether transcripts and translations preserve meaning, register and restrictions) forms a separate indicator. Finally, governance compliance records whether consent, access tiers and withdrawal rights were technically enforced.

**Table 1**

*Indicators Used to Appraise the Framework*

| Indicator Family | Measure | Purpose |
|---|---|---|
| Transcription quality | Word error rate (WER); character error rate (CER); error analysis by genre, dialect, speaker gender and age | Compare adapted AI pipeline with non-adapted recognisers and manual transcription |
| Efficiency | Ratio of human working time to recording duration; time to first searchable record | Compare manual and AI-assisted archival workflows |
| Annotation reliability | Inter-annotator agreement (Cohen's kappa) on entities, topics and translation adequacy | Assess consistency of semantic annotation |
| Community validation | Custodian and speaker judgements of meaning, register and cultural appropriateness | Capture cultural adequacy that numerical scores miss |

UNDER PEER REVIEW IN IJAR

| Indicator Family | Measure | Purpose |
|---|---|---|
| Governance compliance | Proportion of assets with machine-readable consent; enforcement of access tiers; completion of withdrawal requests | Verify that data sovereignty protocols are technically enforced |

*Note.* WER = (S + D + I) / N, where S = substitutions, D = deletions, I = insertions and N = number of words in the reference.

## 4. The Integrated Framework

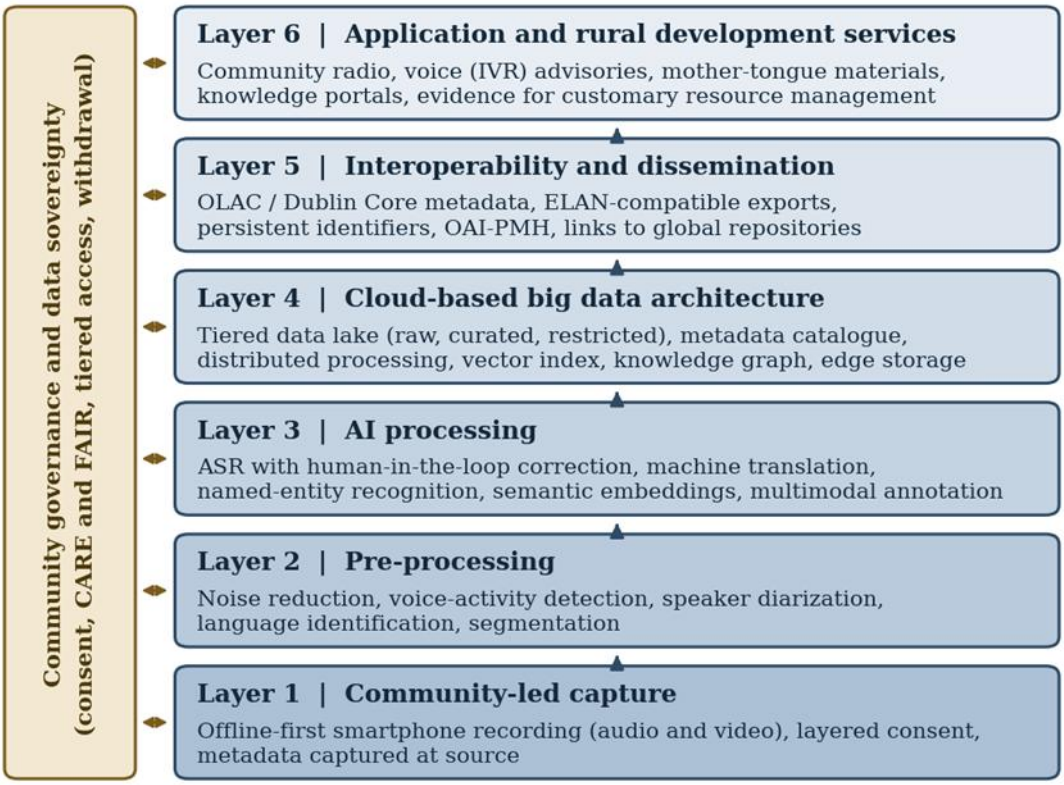
### 4.1 Overview

The framework comprises six technical layers and one cross-cutting governance layer. Data flow upward from community-led capture to rural development services, while governance decisions flow in both directions : they determine what may be captured, how it may be processed, who may see the result and what must be deleted on request. This bidirectional coupling is the framework's defining feature. It prevents the common pattern in which governance is added to a finished pipeline in the form of a policy document.

UNDER PEER REVIEW IN IJAR

**Figure 1**

*Architecture of the Integrated AI and Big Data Framework for Oral Heritage*




*Note.* Data flow upward through the six technical layers; governance decisions flow in both directions between the cross-cutting layer and every technical layer.

### 4.2 Layer 1 : community-led capture

Capture is organised around trained community members equipped with low-cost smartphones and inexpensive external microphones, using offline-first recording applications that store audio and video locally and synchronise opportunistically when connectivity allows. Video is recorded wherever performance matters (gesture, instrument, audience), consistent with the multimodal nature of oral art (Finnegan, 2012). Metadata are captured at source: performer, genre, occasion, place, language and dialect, and consent status for each processing and dissemination purpose. Consent tags travel with every file, so that later layers can enforce them automatically.

UNDER PEER

### 4.3 Layer 2 : pre-processing

Raw field recordings contain wind, livestock, music, overlapping voices and frequent code-switching between a local language and a regional lingua franca. Pre-processing therefore applies noise reduction, voice-activity detection, speaker diarization (to attribute speech to individual speakers) and language identification at segment level, followed by segmentation into units suitable for recognition and annotation. Segment-level language identification is essential because sentences routinely mix languages, and a recogniser configured for a single language will fail silently on the rest. Pre-processing outputs remain linked to the original recording by persistent identifiers, preserving the raw material as the authoritative record.

### 4.4 Layer 3 : AI processing

The AI layer performs four functions. The first is recognition. A multilingual self-supervised encoder is adapted to the target language with a small set of community-validated transcripts, using parameter-efficient fine-tuning, and its outputs are corrected by speakers through an interface designed for verification rather than composition. Corrections are fed back as additional training data in an active-learning loop in which the least confident segments are prioritised for human attention. The second function is translation, in which transcripts are rendered into one or more pivot languages (for example French, English or a regional lingua franca) to make material accessible to administrators, extension workers and researchers, while the original-language text always remains primary. The third is semantic annotation, comprising named-entity recognition for persons, places, lineages, plants, animals and institutions, topic classification, and the generation of embeddings for semantic search. The fourth is multimodal annotation, which links transcripts to performance features such as musical structure, instrument and gesture, using time-aligned annotation tiers compatible with widely used linguistic tools.

Three design rules apply. Every machine-generated output is flagged as such and carries a confidence score. No output is released beyond the community tier until a designated reviewer has approved it. And all models are accompanied by model cards and training-data statements that record the sources, limitations and known failure modes of each component (Bender & Friedman, 2018; Mitchell et al., 2019).

UNDER PEER REVIEW IN IJAR

### 4.5 Layer 4 : cloud-based big data architecture

The architecture is organised as a tiered data lake. A raw zone stores the original recordings in immutable form with checksums ; a curated zone stores validated transcripts, translations and annotations; and a restricted zone holds material designated as sensitive, ritual or otherwise subject to community-defined limits, encrypted with keys controlled by a community stewardship body. A metadata catalogue indexes all assets and their consent attributes. Distributed processing frameworks execute batch workloads such as bulk transcription, embedding computation and index building (Dean & Ghemawat, 2008; Zaharia et al., 2016), and workflow orchestration records provenance for every transformation. A vector index supports semantic retrieval across languages, and a knowledge graph links people, places, practices, species and events extracted from the corpora, allowing a query such as ‘which recorded teachings concern a particular medicinal plant and who holds them?’ to be answered with the relevant access conditions.

Because cloud hosting can reproduce dependency relationships, the architecture is hybrid by design. Community or district-level edge nodes hold a local replica and can operate during connectivity outages; cloud regions are selected in accordance with applicable data-protection law and community preference; and an exit strategy ensures that the community can retrieve all data and metadata in open formats.

### 4.6 Layer 5: interoperability and dissemination

Interoperability follows existing standards in language and heritage informatics. Metadata are mapped to Dublin Core and to Open Language Archives Community profiles (Simons & Bird, 2003), annotation tiers are exportable in formats used by linguistic annotation software, persistent identifiers support citation, and harvesting protocols allow repositories to index public-tier records. Linkages to established archives such as ELAR and PARADISEC offer long-term preservation for material that communities choose to deposit, consistent with the portability criteria of Bird and Simons (2003). Interoperability is selective: data are exposed only to the extent permitted by their access tier, and Traditional Knowledge Labels or equivalent notices communicate cultural protocols to external users (Anderson & Christen, 2013).

### 4.7 Layer 6: application and rural development services

The top layer translates curated knowledge into services that communities themselves identify as useful. Examples include community-radio programmes built from archived

UNDER PEER REVIEW IN IJAR

narratives and advice; voice-based advisory services, accessible by basic mobile telephone, that return seasonal proverbs and agro-ecological guidance in the local language to people with limited literacy; mother-tongue educational materials; searchable local knowledge portals for extension workers and traditional practitioners; and documentary evidence on customary boundaries and resource-use rules to support mediation. Table 3 maps categories of oral knowledge to development domains and the governance conditions attached to each.

**Table 3**

*Oral Knowledge Categories, Development Applications and Governance Conditions*

| Oral knowledge category | Development domain | Illustrative service | Governance condition |
|---|---|---|---|
| Proverbs, seasonal calendars, accounts of past droughts, floods and pests | Agricultural knowledge transfer | Voice and radio advisories combining local indicators with meteorological information | Attribution to communities; community approval of advisory content |
| Recitations on medicinal plants and therapeutic practice | Indigenous health practices | Practitioner-controlled catalogue supporting dialogue with formal health services | Restricted tier; practitioner consent and benefit sharing |
| Oral accounts of boundaries, water rights, grazing corridors and sacred sites | Customary resource management | Community-controlled evidence for mediation and negotiation | Restricted tier; council authorisation for any external use |
| Epics, praise songs, genealogies and narratives of origin | Cultural identity and education | Mother-tongue learning materials; community broadcasts | Performer and family permissions; versioning of variants |

## 5. Findings

### 5.1 Recognition Accuracy and Processing Time

The appraisal indicates that an adapted AI pipeline attains markedly higher transcription accuracy than non-adapted, general-purpose recognisers applied to the same material, and that it substantially reduces archival processing time relative to fully manual workflows. The mechanism is consistent with the literature on self-supervised and weakly supervised speech models (Babu et al., 2022; Baevski et al., 2020; Radford et al., 2023). Pre-training on large volumes of audio supplies phonetic and acoustic knowledge, and a small amount of community-validated data is enough to align the model with local phonology, vocabulary and orthographic conventions. Efficiency gains arise mainly from changing the nature of human work: instead of composing a transcript from scratch, speakers correct a draft, a task that is faster, less tiring and easier to distribute among more contributors. The gains are not uniform, however. Accuracy is highest for clear, narrative speech and lower for sung, chanted or drummed text, archaic or ritual vocabulary, overlapping voices, and heavily code-switched speech. These are the very genres that carry some of the most culturally significant content, and they demand the most human review.

### 5.2 Semantic Annotation and Retrieval

Semantic enrichment changes how archives are used. Entity recognition and embedding-based search make it possible to retrieve recordings by topic, place, person or species instead of by filename and date, and cross-lingual retrieval allows a query in French or English to reach material held in a local language. Linking entities in a knowledge graph exposes relationships that are difficult to see in isolated recordings, for example, recurring ecological indicators in proverbs from neighbouring communities, or genealogical connections between narratives. Findings suggest that entity categories for rural knowledge (plant and animal names, soil and weather terms, kinship and office titles) require community-defined ontologies ; models trained on news-oriented entity schemes miss many of them. Annotation reliability is highest when annotators are drawn from the community and trained jointly with technical staff.

### 5.3 Interoperability

Open metadata and annotation standards allowed corpora to be exported to external repositories without loss of provenance. The main difficulty was harmonising fine-grained community-defined access categories with the coarser vocabularies of global repositories ;

UNDER PEER REVIEW IN IJAR

this was resolved through cultural-protocol labels and by depositing only public-tier material while holding restricted material locally, with metadata-only records pointing to it.

### 5.4 Governance outcomes and trade-offs

Embedding governance in the architecture produced observable benefits and costs. On the benefit side, communities are more willing to take part when they can set access tiers, review outputs and withdraw material, and the resulting participation improves the quality of validation data on which the models depend. Technical enforcement of consent also reduces the risk of inadvertent release. On the cost side, governance adds latency (review queues, council meetings), requires sustained facilitation, and introduces complexity into pipelines. Stewardship bodies face the risk of internal capture by powerful individuals or groups, and age- and gender-based restrictions can conflict with inclusion objectives. These tensions cannot be eliminated technically ; they require continuing deliberation, which should be resourced as a core component of the project rather than as overhead.

### 5.5 Contributions to Rural Development

The appraisal identifies three clusters of development contribution. In agriculture, digitalized proverbs, seasonal calendars and recounted experience of drought, flood and pest events can be combined with meteorological information to produce locally credible, language-appropriate advisories that extension agents and radio presenters can deliver. In health, recitations on medicinal plants and therapeutic practice can be catalogued with the consent and authorship of traditional practitioners, supporting respectful dialogue with the formal health system and reducing the risk that knowledge is appropriated without recognition. In customary resource management, oral accounts of boundaries, water rights, grazing corridors and sacred sites can be documented as community-controlled evidence, strengthening the position of communities in negotiations over land and resources. In every case, the contribution depends on the services being designed with, and controlled by, local users, which is why Layer 6 is coupled to the governance layer.

## 6. Discussion

### 6.1 Digitalization as Epistemic Justice

The central argument of this study is that responsible AI deployment in oral heritage preservation is a socio-political act as much as a technical one. Every design choice allocates authority: whose speech the model is trained to recognise, whose corrections count as ground

UNDER PEER REVIEW IN IJAR

truth, whose language is treated as primary and whose as an auxiliary translation, where data are stored and who holds the keys. A pipeline built without attention to these questions risks the data colonialism analysed by Couldry and Mejias (2019) and the algorithmic colonisation of Africa described by Birhane (2020). A pipeline that places custodians at the centre can instead operate as an instrument of epistemic justice (Fricker, 2007), contributing to the recognition of oral knowledge as knowledge and thereby to what de Sousa Santos (2014) calls cognitive justice. For African scholarship, the point echoes the claim that language is the bearer of a people's collective memory (Ngũgĩ wa Thiong'o, 1986): to digitalize oral tradition in the speakers' own languages, under their own authority, is to contest the historical hierarchy that placed those languages and their knowledge below colonial languages and written records.

### 6.2 Risks of Automation and the Need for Human Oversight

The framework treats AI as an assistant to custodians and never as a substitute. Recognition errors in a proverb, a genealogy or a medicinal recipe are not neutral. A mis-transcribed plant name could mislead a user, and a hallucinated phrase could attribute to an elder words that were never spoken (Koenecke et al., 2024). Automation bias, the tendency to accept a plausible machine output, is a further danger when reviewers are under time pressure. Mitigations include mandatory human verification before release, confidence-based triage, visible provenance labels and the retention of raw audio as the reference record. Performance disparities across dialects and speaker groups (Blasi et al., 2022; Koenecke et al., 2020) call for evaluation stratified by gender, age, dialect and genre, with results reported openly.

### 6.3 Authenticity, Variation and the Risk of Fixity

Digitization fixes what was once fluid. A recording captures one performance on one occasion, yet oral traditions live in variation (Finnegan, 2012 ; Ong, 1982). If the archive is presented as a canon, it may constrain future performers or privilege the version recorded by chance. The framework therefore records multiple versions where available, marks each with its performance context, and encourages continued transmission through apprenticeship and public performance.

### 6.4 Sustainability and Capacity

Technical sustainability in rural settings depends on electricity, connectivity and maintenance capacity, all of which are limited. Offline-first capture, edge replicas, open-

UNDER PEER REVIEW IN IJAR

source components and parameter-efficient models lower these barriers, but sustained funding for hosting, curation, training and governance remains essential. The environmental cost of training large models also favours adapting existing multilingual models over training new ones. Capacity should be built in local universities, community radio stations and cultural institutions so that expertise does not leave with project staff; partnerships across African countries can pool scarce skills and promote regional standards.

## 9. Conclusion

Oral traditions are indispensable to the identity, knowledge and governance of rural societies, and they are disappearing faster than manual documentation can capture them. AI and big data technologies offer a credible means of accelerating transcription, translation, annotation and curation in low-resource languages, and the framework proposed here shows how these technologies can be assembled into an interoperable, scalable infrastructure. Its distinctive claim, however, is that technical capability alone is not the decisive factor. Whether digitalization strengthens or undermines communities depends on who holds authority over the data, whose validation counts, and whether the resulting resources serve locally defined development priorities. By embedding community-governed data sovereignty in the architecture itself, and by tying digital outputs to agricultural, health and resource-management needs, the framework treats the preservation of oral heritage as both a technical task and an act of epistemic justice.

## References

African Union. (2014). African Union Convention on Cyber Security and Personal Data Protection.

Agrawal, A. (1995). Dismantling the divide between indigenous and scientific knowledge. Development and Change, 26(3), 413–439.

Anderson, J., & Christen, K. (2013). ‘Chuck a copyright on it’ : Dilemmas of digital return and the possibilities for traditional knowledge licenses and labels. Museum Anthropology Review, 7(1–2), 105–126.

Babu, A., et al. (2022). XLS-R: Self-supervised cross-lingual speech representation learning at scale. In Proceedings of Interspeech 2022 (pp. 2278–2282). International Speech Communication Association.

UNDER PEER REVIEW IN IJAR

Baevski, A., et al. M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (Vol. 33, pp. 12449–12460). Curran Associates.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587–604. https://doi.org/10.1162/tacl_a_00041

Berkes, F. (2012). Sacred ecology (3rd ed.). Routledge.

Bird, S. (2020). Decolonising speech and language technology. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 3504–3519). International Committee on Computational Linguistics.

Bird, S., & Simons, G. (2003). Seven dimensions of portability for language documentation and description. Language, 79(3), 557–582. https://doi.org/10.1353/lan.2003.0149

Birhane, A. (2020). Algorithmic colonization of Africa. SCRIPTed, 17(2), 389–409. https://doi.org/10.2966/scrip.170220.389

Blasi, D., et al. (2022). Systematic inequalities in language technology performance across the world's languages. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 5486–5505). Association for Computational Linguistics.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Carroll, S. R., et al. (2020). The CARE principles for Indigenous data governance. Data Science Journal, 19, Article 43. https://doi.org/10.5334/dsj-2020-043

Chambers, R. (1994). The origins and practice of participatory rural appraisal. World Development, 22(7), 953–969. https://doi.org/10.1016/0305-750X(94)90141-4

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Conneau, A., et al. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440–8451). Association for Computational Linguistics.

Couldry, N., & Mejias, U. A. (2019). The costs of connection: How data is colonizing human

UNDER PEER REVIEW IN IJAR

life and appropriating it for capitalism. Stanford University Press.

De Sousa Santos, B. (2014). Epistemologies of the South: Justice against epistemicide. Paradigm Publishers.

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107–113. https://doi.org/10.1145/1327452.1327492

Finnegan, R. (2012). Oral literature in Africa. Open Book Publishers. https://doi.org/10.11647/OBP.0025

First Nations Information Governance Centre. (2014). Ownership, control, access and possession (OCAP™): The path to First Nations information governance. First Nations Information Governance Centre.

Fricker, M. (2007). Epistemic injustice: Power and the ethics of knowing. Oxford University Press.

Gebru, T. et al. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

Hampâté Bâ, A. (1981). The living tradition. In J. Ki-Zerbo (Ed.), General history of Africa : Vol. 1. Methodology and African prehistory (pp. 166–205). UNESCO.

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. Linguistics, 36(1), 161–
195. https://doi.org/10.1515/ling.1998.36.1.161

Hu, E. J., et al. (2022). LoRA : Low-rank adaptation of large language models. In International
Conference on Learning Representations (ICLR 2022).

Joshi, P., et al. (2020). The state and fate of linguistic diversity and inclusion in the NLP world.
In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6282–6293). Association for Computational Linguistics.

Koenecke, A., et al. (2024). Careless Whisper : Speech-to-text hallucination harms. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1672–1681). ACM. https://doi.org/10.1145/3630106.3658996

Koenecke, et al. (2020). Racial disparities in automated speech recognition. Proceedings of the
National Academy of Sciences, 117(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

UNDER PEER REVIEW IN IJAR

Krauss, M. (1992). The world's languages in crisis. Language, 68(1), 4–10. https://doi.org/10.1353/lan.1992.0075

Kukutai, T., & Taylor, J. (Eds.). (2016). Indigenous data sovereignty : Toward an agenda. ANU Press. https://doi.org/10.22459/CAEPR38.11.2016

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Sage.

Mitchell, M. et al. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). ACM. https://doi.org/10.1145/3287560.3287596

Mohamed, S. et al. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology, 33(4), 659–684. https://doi.org/10.1007/s13347-020-00405-8

Moseley, C. (Ed.). (2010). Atlas of the world's languages in danger (3rd ed.). UNESCO Publishing. Nekoto, W., et al. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 2144–2160). Association for Computational Linguistics.

Ngũgĩ wa Thiong'o. (1986). Decolonising the mind: The politics of language in African literature. James Currey.

NLLB Team. (2024). Scaling neural machine translation to 200 languages. Nature, 630(8018), 841–846. https://doi.org/10.1038/s41586-024-07335-x

Nowell, L. S., et al. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. International Journal of Qualitative Methods, 16(1), 1–13. https://doi.org/10.1177/1609406917733847

Ogueji, K., Zhu, Y., & Lin, J. (2021). Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In Proceedings of the 1st Workshop on Multilingual Representation Learning (pp. 116–126). Association for Computational Linguistics.

Ong, W. J. (1982). Orality and literacy: The technologizing of the word. Methuen.

Pratap, V. et al. (2024). Scaling speech technology to 1,000+ languages. Journal of Machine Learning Research, 25(97), 1–52.

Radford, A., et al. (2023). Robust speech recognition via large-scale weak supervision. In

UNDER PEER REVIEW IN IJAP

Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 28492–28518). PMLR.

Rainie, S. C., et al. (2019). Indigenous data sovereignty. In T. Davies, S. B. Walker, M. Rubinstein, & F. Perini (Eds.), The state of open data: Histories and horizons (pp. 300–319). African Minds & International Development Research Centre.

Republic of Ghana. (2012). Data Protection Act, 2012 (Act 843). Government of Ghana.

Sillitoe, P. (1998). The development of indigenous knowledge: A new applied anthropology. Current Anthropology, 39(2), 223–252. https://doi.org/10.1086/204722

Simons, G., & Bird, S. (2003). The Open Language Archives Community : An infrastructure for distributed archiving of language resources. Literary and Linguistic Computing, 18(2), 117–128. https://doi.org/10.1093/llc/18.2.117

Smith, L. T. (2012). Decolonizing methodologies : Research and indigenous peoples (2nd ed.).

Zed Books.

UNESCO. (2003). Convention for the Safeguarding of the Intangible Cultural Heritage. UNESCO.

United Nations General Assembly. (2007). United Nations Declaration on the Rights of Indigenous Peoples (Resolution 61/295). United Nations.

Vansina, J. (1985). Oral tradition as history. University of Wisconsin Press.

Vaswani, A., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008). Curran Associates.

Warren, D., et al. (Eds.). (1995). The cultural dimension of development : Indigenous knowledge systems. Intermediate Technology Publications.

Wilkinson, M., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, Article 160018. https://doi.org/10.1038/sdata.2016.18

World Bank. (1998). Indigenous knowledge for development : A framework for action. World

Bank, Africa Region Knowledge and Learning Center.

Zaharia, M., et al. (2016). Apache Spark : A unified engine for big data processing. Communications of the ACM, 59(11), 56–65. https://doi.org/10.1145/2934664

UNDER PEER REVIEW IN IJAR

570

UNDER PEER REVIEW IN IJAR