

MLHKP-MCD: Evidence-Preserving Digital Infrastructure for Munda Living Heritage

Mohammad Amir Khusru Akhtar

akakhtar.2024@gmail.com

Usha Martin University

Arvind Hans

Usha Martin University

Rajan Pahan

Usha Martin University

Article

Keywords:

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10949983/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

MLHKP-MCD: Evidence-Preserving Digital Infrastructure for Munda Living Heritage

Mohammad Amir Khusru Akhtar* Arvind Hans Rajan Pahan

Usha Martin University, Ranchi – 834001, Jharkhand, India

*Corresponding author: akakhtar.2024@gmail.com

Abstract

Living-heritage datasets can preserve records while losing the evidential relationships needed to interpret them responsibly. We present the Munda Living Heritage and Knowledge Project and its Munda Cultural Dataset, a provenance-aware digital research infrastructure that keeps sources, exact passages or field segments, evidence, claims, cultural indicators, context, verification, variation, rights, and access conditions distinct but queryably connected. The system combines a domain ontology, permanent identifiers, a federated source census, evidence and contradiction links, culturally constrained access classes, release-indexed completeness audits, and a synchronized research interface. Three implemented algorithms formalize ingestion, evidence-to-claim linking, and release auditing. Repository-generated validation checks integrity and corpus maturity without manually entered counts. The architecture supports historical specificity, uncertainty, disagreement, and community authority as explicit research properties, offering a reproducible framework for evidence-preserving digital heritage and responsible knowledge stewardship.

1 Introduction

A cultural dataset may contain thousands of records and still leave a basic question unanswered: where did this claim come from? For living heritage, that question quickly leads to others. Who reported the statement? Where and when was it observed? Was the wording checked against a scan or inherited from noisy OCR? Does another account disagree? Is the record historical, contemporary, field documented, community validated, or restricted? Can the underlying material be displayed or reused? These details are part of the evidence needed to interpret a record responsibly.

The documentary landscape surrounding Munda language, culture, history, and living heritage makes this problem especially visible. Historical works such as Roy [1] and the multi-volume *Encyclopaedia Mundarica* initiated by Hoffmann and collaborators [2] coexist with later scholarship on mortuary practice [3], Indigenous foods and nutrition [4], ethnobotany [5], and oral or literary tradition [6]. These sources differ in geography, historical period, purpose, method, language, terminology, evidential status, and reuse conditions. Flattening them into a single present-tense factual table would erase precisely the differences that make them scientifically interpretable.

The Munda Living Heritage & Knowledge Project (MLHKP) was built to preserve that chain. Its structured evidence engine, the Munda Cultural Dataset (MCD), keeps the source, evidence,

claim, cultural indicator, context, verification state, and access conditions as separate records and links them so that a reader can move from one to the next. The architecture is summarized in figure 1.

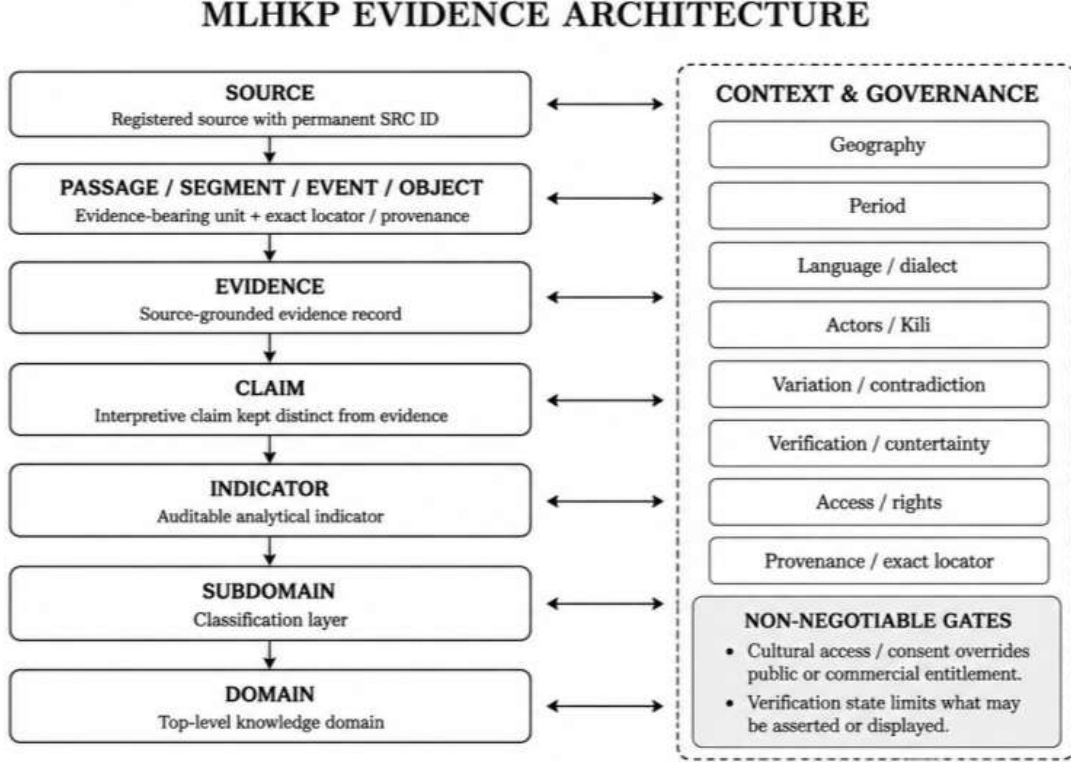


Figure 1: Evidence-preserving architecture of MLHKP-MCD. Context and governance dimensions remain attached across the source-to-domain chain instead of being collapsed into a single factual row.

The paper describes five parts of the implemented system: a Munda-focused information model; an evidence graph that separates source material, evidence, and claims; permanent identifiers and governance fields for access, rights, variation, and uncertainty; reproducible procedures for ingestion, evidence linking, and release auditing; and a release-based evaluation in which reported counts and figures are taken from the checked repository state.

The FAIR principles established findability, accessibility, interoperability, and reusability as central goals for scientific data stewardship [7]. Their emphasis on machine-actionable discovery and reuse is foundational for digital research infrastructure. Indigenous and community-linked data, however, introduce an additional governance problem: technically reusable data may still be culturally inappropriate to expose or repurpose. The CARE principles respond by centering collective benefit, authority to control, responsibility, and ethics [8]. Recent research on digital Indigenous knowledge stewardship similarly argues for relationality, dialog, care, responsibility, and respect rather than treating access as an unconditional objective [9].

MLHKP therefore treats FAIR-like machine readability as necessary but not sufficient. Heritage-science experience shows that linked and reusable data practices must retain historical and cultural context rather than forcing heterogeneous materials into generic scientific templates [10]. A record

can be technically discoverable while remaining culturally restricted; a publicly reachable source can remain unsuitable for redistribution; and a source-reported historical statement can remain unverified as a description of current practice. These distinctions are represented in the data model rather than left to prose documentation.

Mukurtu demonstrated the importance of community protocols, differentiated access, and multiple community records in digital cultural heritage [11, 12]. In the Indian context, open-source Indigenous cultural-heritage systems have likewise been explored as a way to align local needs with digital archive design [13]. Comparative studies of digital language archives further show that preservation, metadata, access, sustainability, and repository architecture must be considered together [14]. Local Contexts extends this governance space through Traditional Knowledge Labels that communicate community-specific provenance, protocol, and permission expectations, including culturally sensitive, secret or sacred, seasonal, gender-restricted, non-commercial, and collaboration-related conditions [15].

Recent work on Indigenous data sovereignty reinforces the need to place governance inside digital infrastructure rather than add it after dissemination. Studies of sovereign data collection and governance emphasize community-determined access, control, trust, and culturally appropriate linkage [16, 17]; work focused specifically on intangible cultural heritage similarly argues for Indigenous data-sovereignty principles by design in digitization arrangements [18]. This orientation is also consistent with UNESCO’s framing of intangible cultural heritage as living knowledge and practice transmitted and recreated by communities [19].

MLHKP does not replace these frameworks. Its complementary contribution is evidential: it asks how community-aware access can coexist with exact source locators, contradictory descriptions, time- and place-specific claims, machine-auditable provenance, and explicit research gaps inside the same research infrastructure.

CIDOC CRM remains a major conceptual standard for integrating heterogeneous cultural-heritage information [20], while the W3C PROV family provides a general model for representing entities, activities, agents, and derivation in interoperable provenance descriptions [21]. Knowledge-graph approaches increasingly support cultural-heritage integration, discovery, narrative construction, and user interaction [22–24]. Recent work also emphasizes end-user curation [25], provenance and traceability [26], and the importance of accessible digital organization of heterogeneous cultural knowledge [27]. A 2026 systematic review of 248 cultural-heritage knowledge-graph studies identifies persistent challenges in ontology alignment, interoperability, evaluation consistency, and methodological transparency [28]. Complementary 2026 work on LLM-assisted heritage-text knowledge graphs shows that rich scholarly discourse requires explicit treatment of evidence, competing hypotheses, uncertainty, and human oversight rather than simple categorical extraction [29].

The distinction introduced here is not that graphs are new, but that evidential status is explicitly modeled across a source-to-claim chain. In MLHKP, a graph relation should not silently transform a historically reported observation into a timeless cultural fact. The system therefore distinguishes evidence-bearing units, interpretive claims, cultural indicators, contextual variation, verification, and access conditions as separate but connected layers.

The functional comparison in table 1 positions MLHKP against influential frameworks and research infrastructures. The approaches are not equivalent products, so the comparison is intentionally limited to design functions rather than overall ranking.

Table 1: Functional design comparison. “Primary” indicates a central purpose of the approach; “supported” indicates an explicit or natural capability; “not primary” indicates that the function lies outside its main design objective.

Approach	Machine reuse	Community authority	Explicit evidence chain	Variation / contradiction	Heritage ontology	Research interface
FAIR CARE Mukurtu protocols	Primary Supported Supported	Not primary Primary Primary	Partial Not primary Partial	Not primary Not primary Supported through community records	Not primary Not primary Platform-specific	Not primary Not primary Primary
Local Contexts CIDOC CRM	Metadata-level Primary	Primary Not primary	Not primary Model-supported Variable	Not primary Model-supported Variable	Not primary Primary	Not primary Not primary
Contemporary heritage KGs MLHKP-MCD	Primary Primary	Variable Primary constraint	Variable Primary	Variable Primary	Primary / variable Munda-focused, extensible	Variable Primary

Taken together, the literature suggests five requirements for a research-grade living-heritage dataset: machine-actionable structure, provenance at the level of specific evidence, explicit representation of interpretive plurality, culturally governed access, and reproducible release evaluation. Existing approaches address important subsets of these requirements. The design objective of MLHKP-MCD is to make all five simultaneously queryable without claiming that a single ontology or platform can resolve community authority, historical interpretation, or rights questions on its own.

2 Methods

2.1 Data model: preserving evidence without flattening culture

The information model is designed around a simple principle: every cultural statement should remain traceable to the best available evidential unit and should carry enough context to prevent accidental universalization. Where applicable, records store permanent identifiers, canonical and alternate names, description, date or period, geography, actors, language or dialect, related entities, source identifiers, exact locator, evidence and verification state, access class, rights or licence status, contributor or reviewer information, uncertainty or confidence, variation and contradiction links, and version timestamps.

The ontology provides explicit homes for people and roles; families, kinship, and Kili; language, alternate forms, meanings, dialects, scripts, pronunciation, morphology, grammar, syntax, usage, and etymology; pregnancy, birth, naming, childhood, youth, courtship, marriage, household, parenthood, ageing, death, funeral, burial, and remembrance; festivals, rituals, ceremonial sequences, prayers, formulae, beliefs, cosmology, and sacred institutions; stories, myths, legends, folk tales, riddles, proverbs, oral histories, songs, Durang, dances, instruments, and performance contexts; food, drink, agriculture, crops, houses, dress, ornaments, tools, crafts, occupations, markets, livelihoods, forests, plants, ethnobotany, animals, seasons, land, customary law, governance, education, health, demography, migration, urbanisation, technology, and media. The same model retains geographic, Kili, family, gender, generational, dialectal, and historical variation rather than forcing a single canonical account.

The ontology is broader than the evidence currently stored in it. A category may have a defined place in the schema even when no evidence has yet been ingested for that category. We therefore report structural coverage and evidential coverage separately.

We built the work in that order. We first defined the evidence model and controlled vocabularies, then implemented the repository schema and interfaces, populated the seed corpus, and ran the integrity checks. The quantitative results reported here were then taken from the checked repository state.

2.2 Source census and corpus construction

The Master Munda Source Census (MMSC) is a federated discovery and registration layer rather than a static bibliography. It records source identity, creator, year, source class, language, geographic scope, canonical or catalogue locator, scan/OCR/full-text state, rights and acquisition status, extraction state, evidence linkage, verification, and provenance. Duplicate digital locators are linked to existing source identities instead of being counted as new intellectual sources.

The ingestion procedure is summarized in algorithm 1. It states the integrity rules used for repository ingestion: validate source and access metadata, preserve permanent identifiers, reject identifier collisions, normalize accepted records, and retain an audit trail. Successful ingestion is an engineering event only; it does not upgrade the cultural truth status of a record.

Algorithm 1. Evidence-preserving source-bundle ingestion

Input: Candidate bundle B and permanent-ID registry R
Output: Accepted canonical bundle with audit report, or explicit rejection
 Validate required envelope fields, record type, source identity, and access class
 Enumerate all permanent identifiers in B
if *an identifier repeats inside B* **then**
 | reject B with a duplicate-ID error
end
if *an identifier already exists in R* **then**
 | reject B with a registry-collision error
end
if *B is historical OCR marked scan-verified but lacks a verified transcription* **then**
 | reject B
end
 Route B to the canonical source-type directory
 Write normalized UTF-8 JSON and content hash
 Register identifiers, source type, access class, path, hash, and ingestion time in R
 Emit a machine-readable import report
Do not infer cultural truth status from successful ingestion alone

2.3 Encyclopaedia Mundarica as a stress test, not the whole system

The sixteen-volume *Encyclopaedia Mundarica* is treated as a first-class historical corpus because it stresses several requirements simultaneously: page accounting, OCR uncertainty, multilingual and comparative-language forms, structural indexing, stories and songs, exact locators, source-era interpretation, and rights assessment. It is nevertheless only one source family inside MLHKP. Books, journal articles, theses, government and archival sources, dictionaries, interviews, oral

histories, observations, events, media, maps, objects, and community validations use the same evidence-preserving envelope.

To keep the paper valid across the life of the project, corpus maturity is represented as a release-indexed sixteen-volume matrix rather than a narrative inventory tied to whichever volumes happen to be processed at publication time. The matrix in figure 2 always displays all sixteen volume slots and is regenerated from release metadata. As later releases advance individual volumes through location, page accounting, OCR or working text, verified transcription, structured extraction, and verified completion, the scientific definition of the figure remains unchanged.

Release-indexed corpus maturity matrix (all sixteen volume slots)

Located / registered	✓	✓	✓	·	·	·	·	·	·	·	·	·	·	·	·	·
Page accounted	✓	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
OCR / working text	✓	✓	✓	·	·	·	·	·	·	·	·	·	·	·	·	·
Verified transcription -	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Structured extraction -	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Verified complete -	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI

Encyclopaedia Mundarica volume

Figure 2: Release-indexed maturity matrix for all sixteen *Encyclopaedia Mundarica* volumes. Filled cells indicate that the corresponding release criterion is satisfied; dots indicate that it is not yet satisfied. The figure is generated from the release metadata rather than manually redrawn.

The authoritative ordering for historical text verification is scan or page image, then verified transcription, then working transcription, then raw OCR. OCR is useful for discovery, but machine readability is not treated as philological verification. This rule protects unusual spellings, Mundari forms, comparative-language material, repetition, and apparent printing irregularities from silent “correction.”

2.4 Evidence graph, variation, and scholarly status

The central unit of scholarly responsibility is the evidential relation, not the database row. A source can contain passages; an interview can contain segments; an event can contain steps; an object or media item can supply material evidence. These evidence-bearing units can support, qualify, or contradict claims. Claims can then map to cultural indicators without erasing provenance or evidential status.

The evidence-linking procedure is formalized in algorithm 2. It preserves exact locators, assigns status from acquisition and verification evidence rather than plausibility, and creates variation or contradiction relations when accounts materially differ.

This separation solves a recurrent problem in cultural databases. A historical ethnographic statement can be accurately transcribed and still be an inadequate description of present practice. Accordingly, MLHKP distinguishes source-reported, historical-source-reported, field-documented, community-validated, contested, superseded-interpretation, and restricted states. The graph can

Algorithm 2. Context-preserving evidence-to-claim linking

Input: Evidence-bearing unit u , proposed claim c , contextual metadata m
Output: Contextualized evidence–claim relation or review flag
Assign permanent identifiers to u , evidence record e , and claim c
Attach exact locator metadata to u (page, passage, timecode, event step, or object reference)
Preserve source wording and provenance separately from interpretive claim text
Assign evidence status from acquisition and verification state
Attach known geography, period, language/dialect, actors, uncertainty, access, and rights metadata
Search for existing claims mapped to the same indicator
if *new evidence materially disagrees with an existing claim* **then**
 | create a variation or contradiction relation; do not overwrite the earlier claim
end
Link $e \rightarrow c \rightarrow \text{indicator} \rightarrow \text{subdomain} \rightarrow \text{domain}$
Require designated scholarly or community review before promotion to a stronger validation state

therefore answer not only “what records mention this practice?” but “which source says this, in what period and place, with what verification and access status, and what other records disagree?”

2.5 Governance, access, and rights

The governance model treats cultural access and legal reuse as related but non-identical dimensions. Records can be open, community-access-only, research-restricted, embargoed, confidential, or not-for-publication. Rights and commercial reuse are tracked separately. A publicly downloadable historical scan is not automatically assumed to permit redistribution, AI training, or commercialization. This distinction follows the spirit of CARE’s authority and responsibility principles [8], community-protocol practice [12], and Local Contexts’ explicit signaling of cultural permissions and restrictions [15].

The rule is strictest for sacred, ceremonial, clan-specific, medicinal, burial-related, confidential, or otherwise sensitive knowledge. Collection does not imply publication permission, and payment or product entitlement can never override cultural access. The same principle applies to contradictory or multiple community accounts: variation is represented, not resolved by administrative convenience.

2.6 Release-indexed completeness

A living cultural infrastructure cannot defensibly define “complete” as “nothing new can ever be discovered.” Instead, MLHKP defines completeness operationally for a frozen release. A release can be described as structurally complete or source-comprehensive only relative to an explicit schema, a documented source-census protocol, declared processing targets, and machine-checkable integrity gates. Residual gaps remain part of the release record.

The release gate is given in algorithm 3. It checks schema representation, census registration, identifier integrity, provenance, verification logic, restricted-data leakage, rights assumptions, data-to-interface mapping, and enumerated evidence gaps.

Algorithm 3. Release completeness, provenance, and safety audit

Input: Frozen release D , schema registry S , source census C , interface registry U
Output: PASS or FAIL with machine-readable residual gaps
Verify every required schema category is represented in S
Verify all sources discovered under the documented census protocol are registered in C
Check duplicate permanent IDs and broken ontology/evidence references
Check every public factual claim has traceable source/evidence provenance
Check verification labels against underlying acquisition state (e.g., OCR $\neq$ scan-verified transcription)
Check restricted, embargoed, confidential, and not-for-publication records cannot leak into public outputs
Check rights and commercial-reuse fields are not inferred from public availability
Check every data type maps to at least one research or public interface module in U
Enumerate unavailable sources, unverified passages, geographic/temporal gaps, and fieldwork gaps
if *any hard integrity or safety condition fails* **then**
 | return FAIL with exact causes
end
else
 | return PASS with residual research gaps explicitly retained
end

Missing evidence is recorded explicitly. An unrepresented geography, a weakly evidenced life stage, an unresolved contradiction, or an unverified historical transcription can therefore guide the next source search, field visit, or community consultation.

2.7 Research interface

The public Streamlit interface was developed while the corpus was being built. This gave us a practical way to test the schema: if a record could not be searched, filtered, contextualized, or followed to its supporting evidence, the underlying model needed revision.

The interface exposes the main research paths implemented in the repository, including hierarchy browsing, source-backed claims, indicators, evidence links, source records, and correction workflows. The separate Mundarica verification view tracks progress across the sixteen-volume corpus and is used as a focused test of page accounting, transcription, extraction, and verification states.

A search result is intended to support drilldown from a human-readable topic to permanent claim identifier, evidence record, exact source locator, verification state, access class, and rights status. This responds to the broader cultural-heritage need for interfaces that help users navigate complex, multilayered knowledge rather than merely exposing graph structure or raw records [25, 27].

3 Results

3.1 Quantitative validation and reproducibility

The quantitative results were taken from the public repository state used for this revision. Repository records contain 24 top-level domains, 266 subdomains, 798 candidate indicators, 14 seed sources,

38 source-backed claims, 38 evidence records, and 38 evidence links. The structural and evidential profile is shown in figure 3. These are properties of the released dataset, not measurements of cultural completeness or representativeness.

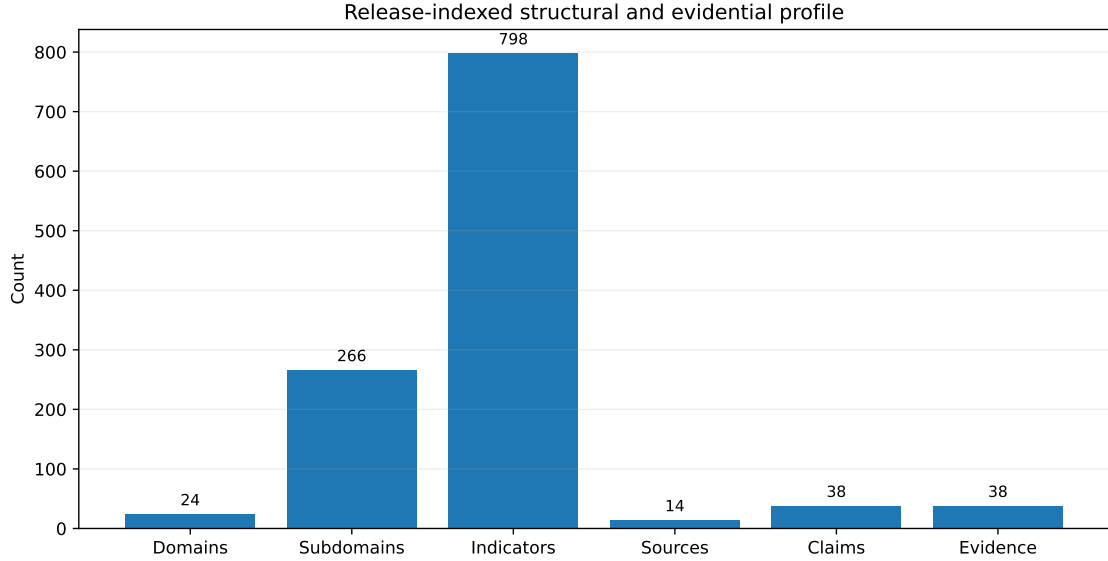


Figure 3: Release-indexed structural and evidential profile. Ontology breadth and evidential population are shown separately so that a large indicator scaffold cannot be mistaken for evidence saturation.

The same repository-backed metrics are reported in table 2. Later releases may change these counts, but the definitions used to compute them remain the same.

Table 2: Audited release metrics. The table is populated from the machine-readable release snapshot used to generate the manuscript assets.

Metric	Release value
Top-level cultural domains	24
Cultural subdomains	266
Adaptive candidate indicators	798
Canonical seed sources	14
Source-backed seed claims	38
Evidence records	38
Evidence links	38

Beyond counts, we checked the repository relations connecting claims, evidence, and evidence links. Every seed claim in the released corpus has a corresponding evidence record and an explicit **supported.by** link. The integrity indicators are visualized in figure 4. They report coverage of these implemented relations rather than a cultural-quality score.

The corresponding numerical audit summary appears in table 3. The 100% values mean only that the released seed claims satisfy the stated repository relations. They do not indicate cultural completeness, correctness of every source interpretation, or community validation.

Reproducibility rests on the repository itself: the data tables, SQLite database, schema, software, documentation, and seed tests are public, while the manuscript package contains the macros and figure-generation assets used for this revision. A later release can therefore be re-audited against

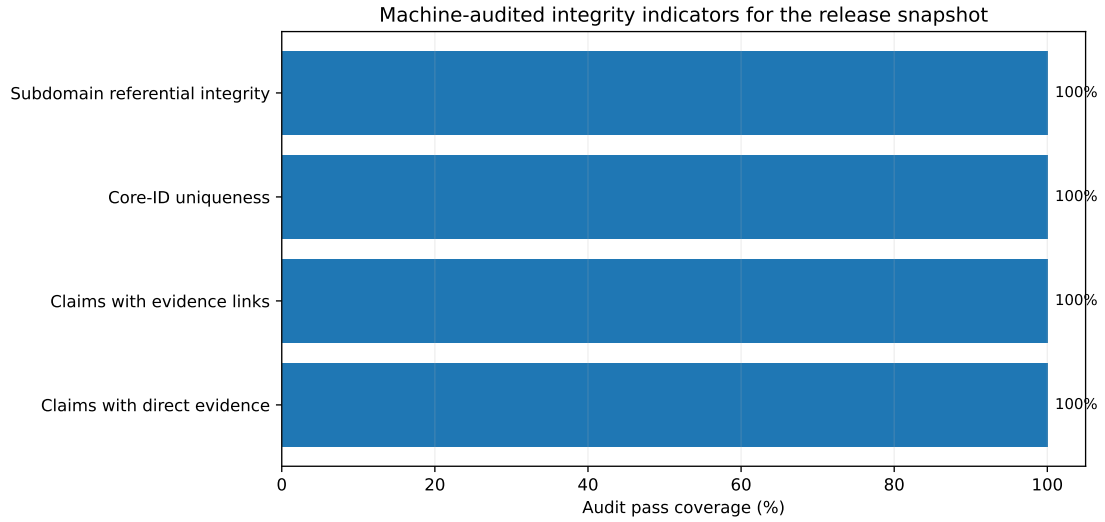


Figure 4: Machine-audited integrity indicators for the release snapshot. The plotted values report the coverage of implemented checks: direct evidence for seed claims, evidence-link coverage, core-ID uniqueness, and subdomain referential integrity.

Table 3: Machine-checkable validation results for the audited seed layer.

Validation condition	Result
Seed claims with direct evidence	100%
Seed claims represented in evidence links	100%
Duplicate core identifiers detected	0
Orphan subdomains detected	0
OCR treated as equivalent to scan-verified transcription	No
Public availability treated as proof of redistribution rights	No

the same definitions without treating the present counts as permanent.

3.2 End-to-end source-to-claim traceability

A direct repository trace illustrates the central provenance claim. Record CLM-000025 is a source-reported claim on Pathalgarri from source SRC-000007, mapped to domain DOM-21 and subdomain DOM-21-S07; its evidence record is EVD-000025, and LNK-000025 explicitly links the claim to that evidence as `supported_by`. The claim is geographically scoped to Ithe, Murhu, Khunti and remains marked `not_field_verified`. This trace is important because the system preserves both the scholarly assertion and its limits: the record can be retrieved as evidence from a named source without being promoted to a universal or contemporary statement about all Munda communities. Figure 5 summarizes this repository-backed path.

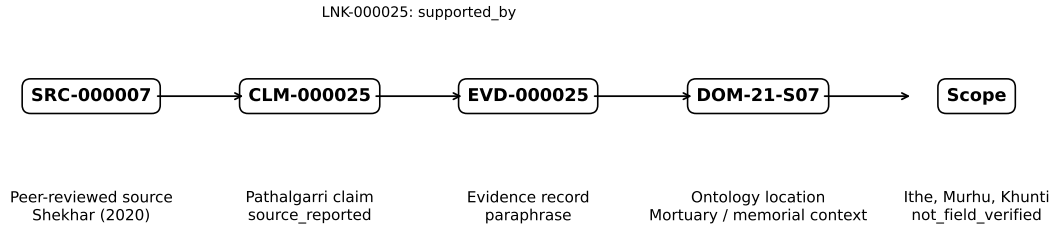


Figure 5: Repository-backed source-to-claim trace for one non-sensitive seed record. The permanent identifiers preserve the path from registered source to claim, evidence, ontology location, geographic scope, and verification state.

The reverse lookup is equally explicit: starting from the claim or ontology node, the repository relations recover EVD-000025 and the originating SRC-000007. This provides a concrete implementation check of source-to-claim traceability in the released repository.

3.3 Research productivity and reuse paths

The practical value of an evidence-preserving dataset is measured by the research work it enables, not only by the number of stored records. MLHKP-MCD is designed so that a question can be transformed into an auditable evidence packet: relevant ontology nodes identify the search space; source and field records supply candidate evidence; exact locators preserve traceability; variation and contradiction links prevent premature consensus; governance filters determine what may be returned; and release metadata records the state of the result.

The principal research workflows are summarized in table 4. Each workflow uses the same underlying evidence graph but produces a different scholarly product, allowing researchers from linguistics, anthropology, history, digital humanities, heritage studies, and information science to reuse the infrastructure without flattening disciplinary distinctions.

Table 4: Representative research workflows enabled by the evidence-preserving model.

Research task	Evidence traversal	Reusable scholarly output
Historical change	Compare time-stamped claims, source passages, places, and validation states for the same indicator	Chronology with source-specific continuities, changes, and disagreements
Language and lexicon	Traverse terms, variants, dialect/geography fields, examples, source locators, and historical forms	Provenance-aware lexical inventory or dialect comparison
Life-course and ritual study	Join indicators with actors, Kili/kinship, event steps, objects, songs, places, and period	Evidence matrix for birth, marriage, death, burial, or festival research
Contradiction analysis	Retrieve claims mapped to the same indicator and inspect supporting or conflicting evidence	Review set distinguishing variation, contradiction, uncertainty, and superseded interpretation
Source-census research	Link source identities, editions/locators, acquisition state, domains, and evidence coverage	Reproducible bibliography, acquisition map, and corpus-coverage audit
Governance-aware reuse	Apply access, consent, rights, and commercial-reuse status before export or reporting	Permitted research bundle with explicit exclusions and reuse conditions

A query on marriage, for example, can combine terminology, historical descriptions, ritual steps, kinship and Kili relations, places, objects, songs, field evidence, temporal change, contradictory accounts, and validation states while preserving the provenance of each assertion. Similar workflows can support historical change in mortuary practice, lexical variation across dialects, transformations in Indigenous food systems, ethnobotanical continuity, or comparative study of ritual vocabulary.

The same evidence graph can also support computational work while keeping the status of each record visible. Retrieval systems can rank passages and still return exact locators and verification states. Contradiction analysis can flag claims for review without deciding in advance which source is correct. Completeness maps can guide source discovery or fieldwork, and reports can be generated while restricted material remains filtered. These uses are consistent with current heritage-KG research on structured scholarly evidence and interpretive disagreement [29], FAIR/linked heritage-data practice [10], and broader graph-based cultural-heritage infrastructure [24, 28].

4 Discussion

The present release shows that a working cultural-heritage repository can keep provenance, interpretation, and access visible at the same time. A statement may be digitally accessible and still be historical, unverified in the field, or restricted from redistribution. For that reason, the repository keeps source identity, claim status, evidence links, geography, verification, and access information available for inspection.

The counts describe the present release, not Munda culture as a whole. The 24 domains, 266 subdomains, and 798 candidate indicators show the breadth of the information model. The 14 seed sources and 38 source-backed claims show how much evidence is currently populated. Complete claim-to-evidence linkage inside this seed layer is a database-integrity result; it is not a claim that Munda living heritage has been comprehensively documented.

The repository-backed trace in figure 5 provides a more direct test of the central design claim. A reader can move from a named source to a bounded cultural claim, its evidence record, ontology location, geographic scope, and verification state, and can reverse that path without discarding the record’s uncertainty. This differs from using provenance only as descriptive metadata attached after extraction. Here, provenance and status are part of the data relation that determines how the record should be interpreted and reused.

CARE, Mukurtu, Local Contexts, and recent work on Indigenous data sovereignty show why technical openness is not enough for Indigenous cultural data [8, 12, 15–18]. The software can record access and rights restrictions, but it cannot grant community permission. That authority remains with the relevant people and communities, and the repository is designed to record and respect those decisions.

Several limitations are important. First, the released corpus is a seed collection, not a representative census of Munda cultural life across regions, generations, dialects, genders, or historical periods. Second, the 798 indicators are candidate structural categories and should not be interpreted as 798 validated cultural findings. Third, most seed claims are source-reported and not field verified; community validation remains incomplete. Fourth, the present evaluation primarily tests repository structure, traceability, and linkage coverage. It does not yet establish retrieval quality, comparative usability, long-term preservation performance, or superiority over

other heritage platforms on the same corpus. Fifth, some source records point to historical or third-party materials that cannot be redistributed, so reproducibility may require consulting the cited source rather than downloading the underlying document from the repository. Finally, the repository currently contains a modest number of seed sources and claims, and additional ingestion may expose ontology or governance cases not represented in this release.

The next step is to test these areas directly. Useful evaluations include expert checks of provenance recovery, task-based usability studies, comparison with another heritage platform using the same corpus, broader community validation, and saturation analyses by region and cultural domain. These can be added without changing the stored source, evidence, or status history.

MLHKP-MCD provides a reproducible way to keep cultural claims connected to their evidence and to the conditions under which they can be interpreted. The present release is small, but the traceability, uncertainty, variation, and access conditions remain visible as the collection grows.

Data Availability

The data, schemas, documentation, seed corpus, and reproducibility materials supporting this study are available in the public repository at https://github.com/Arithmetic-Power-Geometry/Munda_Cultural_Dataset_v1.0. The repository state checked for this revision is the public `main` branch at commit `6b7728aa1209d6723696844faf3d7fa8e330527a`. Third-party source materials are not redistributed when copyright, consent, cultural governance, or reuse conditions do not permit redistribution.

Code Availability

The Streamlit interface, FastAPI service, database schema, utility code, and seed tests are available in the same public repository: https://github.com/Arithmetic-Power-Geometry/Munda_Cultural_Dataset_v1.0. Original code, schemas, documentation, and original dataset compilation are released under Apache-2.0; third-party materials retain their original rights and access conditions.

Ethics and Cultural Governance Statement

MLHKP separates collection from permission to publish. Sacred, ceremonial, clan-specific, medicinal, burial-related, confidential, embargoed, and other culturally sensitive materials can be assigned restricted access. Historical source statements remain historical or source-reported evidence unless stronger contemporary validation is available. Field and community-contributed material requires appropriate consent and governance review before public release. Cultural access restrictions take precedence over product entitlement or payment status.

Author Contributions

Mohammad Amir Khusru Akhtar: conceptualization, methodology, software architecture, data model, research design, formal analysis, writing—original draft, writing—review and editing, and

project leadership. Arvind Hans: project direction, fieldwork planning, source development, community and implementation coordination, writing-review and editing, and project administration. Rajan Pahan: community consultation coordination, meetings and field logistics, culturally appropriate introductions and participant liaison, field-location coordination, and implementation support.

Competing Interests

The authors declare no competing interests.

References

- [1] Roy, S. C. *The Mundas and Their Country*. Calcutta: City Book Society, 1912.
- [2] Hoffmann, J. and Van Emelen, A. *Encyclopaedia Mundarica, Volume I*. Patna: Superintendent, Government Printing, Bihar, 1950.
- [3] Shekhar, H. “An Ethnoarchaeological Study of Mortuary Practices and Megalithic Traditions of the Mundas of Jharkhand”. In: *Man and Environment* 45.1 (2020), pp. 1–12.
- [4] Ghosh-Jerath, S., Kapoor, R., Bandhu, A., Singh, A., Downs, S., and Fanzo, J. “Indigenous Foods to Address Malnutrition: An Inquiry into the Diets and Nutritional Status of Women in the Indigenous Community of Munda Tribes of Jharkhand, India”. In: *Current Developments in Nutrition* 6.9 (2022), nzac102. DOI: 10.1093/cdn/nzac102.
- [5] Islam, A. T. M. R., Hasan, M. M., Islam, M. T., and Tanaka, N. “Ethnobotanical Study of Plants Used by the Munda Ethnic Group Living around the Sundarbans, the World’s Largest Mangrove Forest in Southwestern Bangladesh”. In: *Journal of Ethnopharmacology* 285 (2022), p. 114853. DOI: 10.1016/j.jep.2021.114853.
- [6] Munda, A. “Study of Literary of Munda Folk Tradition”. In: *The Asian Man* 17.1–2 Special (2023), pp. 137–141. DOI: 10.5958/0975-6884.2023.00017.8.
- [7] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship”. In: *Scientific Data* 3 (2016), p. 160018. DOI: 10.1038/sdata.2016.18.
- [8] Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., and Hudson, M. “The CARE Principles for Indigenous Data Governance”. In: *Data Science Journal* 19 (2020), p. 43. DOI: 10.5334/dsj-2020-043.
- [9] Liew, C. L. and Lipscombe, A. “Centering Dialog and Care in Digital Indigenous Knowledge Stewardship: Of Relationality, Responsibility, and Respect”. In: *Journal of the Association for Information Science and Technology* 75.6 (2024), pp. 671–685. DOI: 10.1002/asi.24873.
- [10] Larsson, Å. M., Bornsäter, B., and Hacke, M. “Developing Practices for FAIR and Linked Data in Heritage Science”. In: *npj Heritage Science* 13 (2025), p. 53. DOI: 10.1038/s40494-025-01598-x.

- [11] Christen, K. A. “Does Information Really Want to Be Free? Indigenous Knowledge Systems and the Question of Openness”. In: *International Journal of Communication* 6 (2012), pp. 2870–2893. URL: <https://ijoc.org/index.php/ijoc/article/view/1618>.
- [12] Christen, K. A., Merrill, A., and Wynne, M. “A Community of Relations: Mukurtu Hubs and Spokes”. In: *D-Lib Magazine* 23.5/6 (2017). DOI: 10.1045/may2017-christen.
- [13] Dutta, A. and Mukhopadhyay, P. “Indigenous Cultural-Heritage Resource Management System: Designing an Open-Source Based Technical Framework”. In: *Journal of Information and Knowledge* 59.2 (2022), pp. 65–77. DOI: 10.17821/srels/2022/v59i2/168787.
- [14] Bharti, S. and Singh, R. K. “Evaluation and Analysis of Digital Language Archives Development Platforms: A Parametric Approach”. In: *The Electronic Library* 40.5 (2022), pp. 552–567. DOI: 10.1108/EL-01-2022-0011.
- [15] Local Contexts. *Traditional Knowledge Labels*. 2026. URL: <https://localcontexts.org/labels/traditional-knowledge-labels/> (visited on 09/06/2026).
- [16] Walker, J. and Kutay, C. “Indigenous Data Sovereignty for Yinhamah”. In: *International Journal on Digital Libraries* 26 (2025), p. 14. DOI: 10.1007/s00799-025-00420-0.
- [17] Marriott, R., Farrell, R., Coffin, J., Reibel, T., and Walker, R. “Recognising Indigenous Data Sovereignty and Implementing Indigenous Data Governance at the Ngangk Yira Institute for Change”. In: *First Nations Health and Wellbeing – The Lowitja Journal* 3 (2025), p. 100072. DOI: 10.1016/j.fnhli.2025.100072.
- [18] Spano, I. and Zhang, Y. “Indigenous Data Sovereignty in Intangible Cultural Heritage Governance: A Complementary Approach to Public–Private Partnerships”. In: *International Journal of Cultural Property* 32.2 (2025), pp. 167–193. DOI: 10.1017/S0940739125100064.
- [19] UNESCO. *Convention for the Safeguarding of the Intangible Cultural Heritage*. 2003. URL: <https://ich.unesco.org/en/convention> (visited on 09/06/2026).
- [20] CIDOC CRM Special Interest Group. *Definition of the CIDOC Conceptual Reference Model, Version 7.1.3*. 2024. URL: <https://cidoc-crm.org/get-last-official-release> (visited on 09/06/2026).
- [21] W3C Provenance Working Group. *PROV-O: The PROV Ontology*. 2013. URL: <https://www.w3.org/TR/prov-o/> (visited on 09/06/2026).
- [22] Carriero, V. A., Gangemi, A., Mancinelli, M. L., Marinucci, L., Nuzzolese, A. G., Presutti, V., and Veninata, C. “ArCo: The Italian Cultural Heritage Knowledge Graph”. In: *The Semantic Web – ISWC 2019*. Vol. 11779. Lecture Notes in Computer Science. Springer, 2019, pp. 36–52. DOI: 10.1007/978-3-030-30796-7_3.
- [23] Fan, Z. and Chen, C. “CuPe-KG: Cultural Perspective-Based Knowledge Graph Construction of Tourism Resources via Pretrained Language Models”. In: *Information Processing & Management* 61.3 (2024), p. 103646. DOI: 10.1016/j.ipm.2024.103646.
- [24] Ferilli, S., Bernasconi, E., Di Pierro, D., and Redavid, D. “The GraphBRAIN Framework for Knowledge Graph Management and Its Applications to Cultural Heritage”. In: *Bridging the Gap Between AI and Reality*. Ed. by B. Steffen. Vol. 14129. Lecture Notes in Computer Science. Cham: Springer, 2025, pp. 144–161. DOI: 10.1007/978-3-031-73741-1_10.

- [25] Mulholland, P., Van Kranenburg, P., Carvalho, J., and Daga, E. “Supporting the End-User Curation of Cultural Heritage Knowledge Graphs”. In: *Proceedings of the 35th ACM Conference on Hypertext and Social Media*. 2024, pp. 35–44. DOI: 10.1145/3648188.3675132.
- [26] Dibowski, H. “Full Traceability and Provenance for Knowledge Graphs”. In: *Formal Ontology in Information Systems*. 2024. DOI: 10.3233/FAIA241309.
- [27] Nappi, M. L., Buono, M., Chivăran, C., and Giusto, R. M. “Models and Tools for the Digital Organisation of Knowledge: Accessible and Adaptive Narratives for Cultural Heritage”. In: *Heritage Science* 12 (2024), p. 112. DOI: 10.1186/s40494-024-01219-z.
- [28] Zhu, L., Abdul Rahman, S., and Timan, H. “Knowledge Graph Applications in Cultural Heritage: A ROSES-Based Systematic Review”. In: *Information* 17.3 (2026), p. 269. DOI: 10.3390/info17030269.
- [29] Schimmenti, A., Pasqual, V., Vitali, F., and van Erp, M. “Knowledge Graphs Generation from Cultural Heritage Texts: Combining LLMs and Ontological Engineering for Scholarly Debates”. In: *Journal of Documentation* 82.7 (2026), pp. 206–250. DOI: 10.1108/JD-07-2025-0203.