Coalition on Digital Impact

July 2026

Defining the Minimum Viable Dataset for Cultural & Linguistic Empowerment

MVD Working Group
Final Report

Table of Content

Abstract

A Working Group assembled by the Coalition On Digital Impact (CODI) developed a framework for underrepresented and low-resource languages to participate in artificial intelligence systems. Rather than focusing solely on technical metrics, the report emphasizes that meaningful inclusion requires a balance of adequate linguistic data, culturally grounded representation, and ethical community governance. The framework introduces the AI Readiness Index (ARI) to categorize a language's current digital presence and a Capability Model that defines the data thresholds needed to unlock specific digital functions, ranging from basic searchability to advanced generative AI. Central to this approach is the Cultural and Ethical Gap Analysis, which ensures that communities maintain authority and consent over their data while preventing the erasure of cultural nuances. Ultimately, the MVD provides a community-led roadmap that allows language groups to define their own goals for digital empowerment without succumbing to extractive or exploitative data practices.

CODI Foreword

Artificial intelligence is being adopted at a speed that outpaces any previous technological shifts. It is quickly becoming the interface through which people access information, education, and public services. In parallel, a growing narrative suggests that language barriers in these systems are being solved. Translation tools support more languages than ever, and speech technologies continue to expand.

This narrative is incomplete.

Most current approaches treat language as a technical problem: how to translate, transcribe, or process it more efficiently. These efforts improve how systems handle language, but they do not ensure that languages themselves are meaningfully represented. Translation is often used as a substitute for presence. Where content does not exist, it is generated or translated from high-resource languages. The result may be readable, but it is frequently detached from the cultural context, knowledge systems, and lived experience of the communities who speak the language.

For many languages, even this workaround is unavailable. Thousands remain digitally invisible, with insufficient digital representation to participate meaningfully in emerging AI and digital ecosystems. Without intervention, the systems shaping the future of information, education, and public services will increasingly reflect only the languages that are already well represented online.

The central issue is not only the quantity of language data. It is how that data is created, structured, and governed. Much of the work to date has focused on improving technology. Less attention has been given to defining what data is required for meaningful participation in digital and AI systems, or to enabling communities to build the datasets that represent their own language and culture. Recognizing that there is no shared framework to do this, the Coalition on Digital Impact (CODI) convened a working group to define a Minimum Viable Dataset (MVD) framework to address this gap.

Rather than defining language inclusion in terms of model performance alone, the MVD establishes a baseline for what it means for a language to exist in digital systems in a way that is usable, culturally grounded, and governed with community consent. It sets the stage and provides a path forward for communities, and the institutions that support them, to create datasets that reflect how their language is used in real life: in communication, education, civic participation, and cultural expression.

The framework defines minimum data requirements, introduces a staged model of AI readiness, and specifies how language data can be structured, labeled, and governed so that it can be discovered and used in digital systems. It also recognizes that technical readiness is only one part of the equation. Community authority, cultural context, and ethical governance are equally central.

This work is grounded in a simple premise: languages should not enter AI systems only through translation or external interpretation. They should be represented directly, through datasets that reflect the communities who speak them.

The MVD does not attempt to solve language representation for communities. It provides a framework they can use to do so on their own terms.

1. Introduction

Approximately 7,000 languages exist globally, yet only a few hundred are represented in digital systems. The vast majority of languages lack sufficient resources for advanced Natural Language Processing ("NLP") and Artificial Intelligence ("AI") applications.[1] Translation tools like Google Translate now support more than 130 languages, including an increasing number of African ones.[2] However, voice assistants and conversational AI systems support far fewer. As of early 2026, Google's Gemini supports just over 65 languages, Apple Intelligence about 15 languages, and ChatGPT around 60 languages.[3] At the same time, around 40% of the world's population lacks access to education or digital services in a language that they speak or understand.[4]This imbalance reflects a structural gap in how languages are represented in digital systems. Languages are not excluded because they lack speakers or cultural relevance, but because they lack the data, infrastructure, and governance frameworks required for inclusion in AI systems. CODI's Minimum Viable Dataset ("MVD") framework addresses this gap by defining the minimum requirements needed for a language to participate meaningfully in digital and AI-enabled systems. Even when data exists, it may not be native to the language: the first large Urdu information-retrieval dataset was developed by translating an English dataset rather than building from original Urdu sources (Butt, Varanasi, & Neumann, 2023), risking the loss of cultural nuance and context-specific meaning.

---

[1] Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://aclanthology.org/2020.acl-main.560/

[2] Google (2024). Africa's digital decade: AI upskilling and expanding speech technology. https://blog.google/company-news/inside-google/around-the-globe/google-africa/africas-digital-decade/

[3] XDA Developers (2026). Here's everything Google added to Gemini in January 2026. https://www.xda-developers.com/everything-google-added-to-gemini-january-2026/. Apple Support (2026). How to get Apple Intelligence. https://support.apple.com/en-us/121115.

[4] UNESCO (2025). Languages matter: Global guidance on multilingual education. https://unesdoc.unesco.org/ark:/48223/pf0000392477

Large numbers of speakers do not guarantee digital presence. Over 30 million people speak Moroccan Arabic (Darija), but it lacks standardized orthography or official recognition. Quechua has approximately 8 to 10 million speakers across South America but has a fragmented digital presence despite its large population. In India and Nepal, Bhojpuri is spoken by over 52 million people but has only limited digital tools and resources.[5]

We define a **community** as any group that self-identifies around a shared language and seeks to create or improve its digital representation. Communities may be linked by place, cultural identity, or shared heritage, often combining multiple elements.

A visible digital presence does not guarantee inclusive representation either. Papua New Guinea illustrates the inverse. Tok Pisin, spoken by approximately 4 to 6 million people, has a relatively strong digital footprint thanks to its use in religious texts, government communication, and online content. Alongside Tok Pisin, more than 800 other languages in Papua New Guinea remain largely absent from digital systems, many of them oral-first and at risk of disappearing. Similar patterns exist in the Solomon Islands and elsewhere in the Pacific, where dozens of local vernaculars remain digitally invisible.

One of those 800 languages is the Kafe Dialect of the Eastern Highlands Province, spoken by more than 100,000 people across the Henganofi and Kainantu Districts. The younger generations raised in urban centers, and even in villages situated along the main Highlands Highway like Kugumo Village, can hear and understand Kafe but cannot speak in their own native vernacular. A key contributing factor is the impact of globalization: current generations are exposed to smartphones and conduct their daily communication, whether in school or at home, in Tok Pisin. Against this backdrop, the PNG Summer Institute of Linguistics (SIL) in Kainantu and the Seventh-Day Adventist Church in Henganofi have collaborated to translate the Holy Bible from English into Kafe in Latin script, with audio and voice recordings. The Bible translation mitigates the loss of language by preserving a digital replica of the text in Latin script alongside aligned audio, both of which can be integrated into digital ecosystems.

For the purposes of this framework, a **language** is any distinct oral or written communication system, standardized or dialectal, that a self-identified community uses and wishes to represent digitally. This includes indigenous, minority, and underrepresented languages (those not used in government or digital systems), non-standard varieties, languages that are mainly oral in nature, and those used mainly for ceremonies and cultural practices linked to identity.

Digital inclusion does not happen at the level of countries or even major languages. It happens at the level of specific language communities. Linguistic and cultural boundaries do not always align. Groups that share a language across regions may operate in very different contexts and may have distinct priorities for how their language is represented digitally. Somali speakers in Somalia and Somali diaspora communities in North America may choose to build joint or

[5] See Joshi et al. (2020) and Ethnologue. https://www.ethnologue.com/

separate datasets depending on whether they identify shared or distinct representational needs. Younger speakers using code-mixed varieties may organize separately from older speakers to document their distinct language use.

For this reason, the MVD framework is designed to support community-defined pathways to AI participation, rather than imposing a single model of readiness or progress.

2. The MVD Framework

The MVD framework defines the minimum dataset requirements necessary for languages and communities to participate in today's digital and AI-driven environment. It identifies dataset thresholds, establishes readiness benchmarks for different levels of digital use, and develops practical pathways that enable communities to build and strengthen their linguistic data resources.

The Structural Gap

Many existing benchmarks and datasets assume written, standardized, and institutionally documented language use. As a result, they systematically exclude oral-first languages, dialectal variation, and culturally contextual forms of communication.

At the same time, language data is often collected without clear consent, governance, or benefit to the communities it represents meaning that:

- AI systems fail to support basic participation for many language communities.
- Linguistic and cultural meaning is flattened or distorted to fit dominant technical norms.
- Communities lack agency over how their language is used or commercialized.
- Funders, developers, and policymakers lack shared criteria for readiness, progress, or accountability.

There is currently no framework that defines what a minimum viable dataset for a language looks like in linguistic, cultural, and governance terms.

The Approach

A central premise of the MVD framework is that linguistic data alone is not sufficient for language inclusion in AI systems. For a language to progress in AI capability, it must be supported not only by adequate linguistic data, but also by culturally grounded representation and ethically governed use. Ideally, these three dimensions (linguistic, cultural, and ethical)

should develop together. Gaps in any one dimension can limit functionality, distort meaning, or result in harmful or extractive outcomes.

Rather than optimizing for maximum model performance, the MVD establishes a shared baseline for inclusion, readiness, and governance. The framework is guided by the following design principles:

- **Minimum requirements reduce exclusion:** Clear, tiered minimum requirements will enable more communities to participate in AI systems without meeting unrealistic or extractive data thresholds.
- **Use-case grounding improves relevance:** Designing datasets around human-centered use cases, rather than abstract benchmarks, will produce AI systems that better serve real community needs.
- **Technical and community readiness enables participation:** Assessing readiness along both technical and governance dimensions enables ethical participation without forcing unwanted AI capabilities.
- **Tiered progression supports community agency:** Allowing communities to remain at, advance through, or stop at, specific tiers supports intentional, values-aligned AI engagement.
- **Embedded governance reduces extractive practices:** Explicit consent, ownership, and stewardship requirements promote data practices that are beneficial to source communities.
- **A shared reference improves coordination:** A common MVD framework will better align funders, developers, and policymakers, reducing fragmentation and misaligned expectations.

The responsibility for enabling AI readiness does not rest on language communities alone. It also lies with those who design, fund, and deploy AI systems. Communities are partners and decision-makers.

The Ecosystem of Actors

AI readiness exists within a broader ecosystem. No single entity can fully develop or sustain a language in AI systems. Progress depends on coordination across five groups:

- **Community and Cultural Custodians.** Language elders, storytellers, poets, educators, religious leaders, and community radio practitioners who carry linguistic knowledge and cultural context.
- **Local and Regional Institutions.** Universities, cultural centers, language academies, and community-based organizations that support documentation, research, and education.

- **Youth and Grassroots Language Actors.** Informal language activists, urban hybrid language users, and digital community organizers who contribute to evolving forms of language use.
- **Technical and Implementation Actors.** AI researchers, dataset practitioners, open knowledge platforms, and tool developers who build systems and infrastructure.
- **Ethical, Policy, and Funding Actors.** Indigenous data governance experts, NGOs, digital rights organizations, and mission-aligned funders who enable and regulate language initiatives.

This distribution ensures that language data is treated not as a purely technical resource, but as a cultural and social asset.

The Working Group identified universities as the most likely practical implementers of the framework, in partnership with communities and supporting funders.

The MVD framework is not a benchmark, a single readiness model, or a fixed timeline; it is a structure for communities to make their own decisions about if and how their languages enter AI systems.

3. Three Elements for Language Inclusion in AI Systems

Three key elements of the MVD framework can help to assess a language's readiness to be incorporated into AI systems. They are independent and should not be conflated.

1. The **AI Readiness Index** describes where the language currently is in AI ecosystems.
2. The **AI Capability Model** identifies which AI functions become available as data and systems develop
3. The **Cultural and Ethical Gap Analysis** identifies deficiencies in ethical governance and highlights gaps in cultural richness and understanding.

A language can score well on one and poorly on another and the framework is built around the fact that they do not move in lockstep.

3.1 The AI Readiness Index: Where the Language Is Today

The AI Readiness Index (ARI) is a structured way to assess how a language currently appears in AI systems. Rather than focusing on future potential, it looks at what already exists: the data, the tools, and the digital presence needed to support participation in AI systems and AI-enabled environments.

The ARI is based on three components:

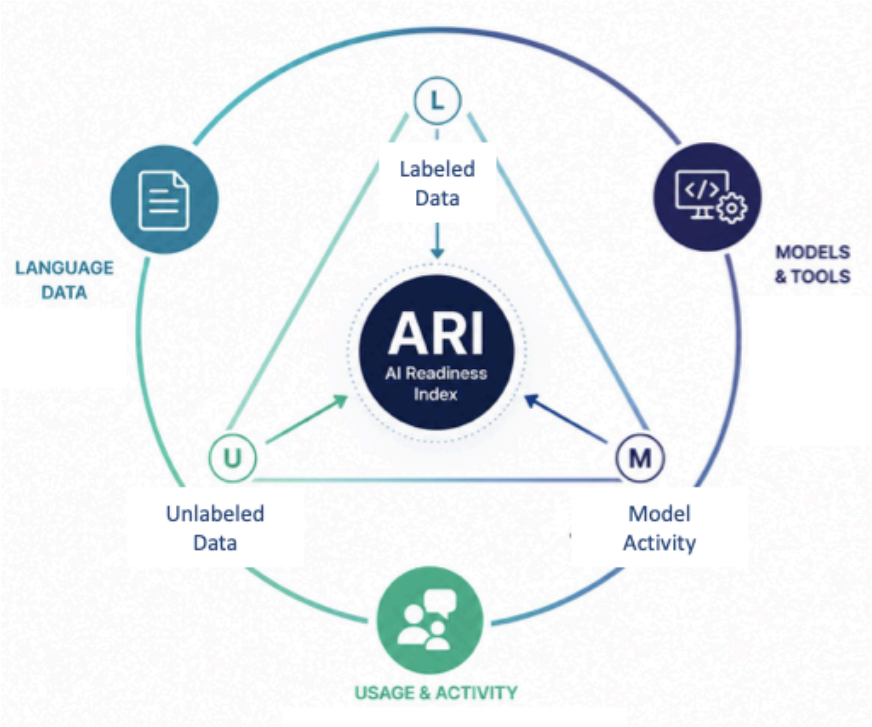
- **Language Data**: the availability of text, speech, or other linguistic data in digital form, including both structured datasets and unstructured sources such as websites or community-generated content.
- **Models and Tools:** whether AI systems have been developed for the language, such as translation tools, speech systems, keyboards, or language models.
- **Usage and Activity:** the extent to which the language is visible and actively used in digital spaces and AI-related contexts.

These are scored through three variables:

- **Labeled Data (L):** structured and annotated datasets used to train AI systems.
- **Model Activity (M)**: the development and deployment of AI tools and models.
- **Unlabeled Data (U):** raw digital language data that reflects broader usage.

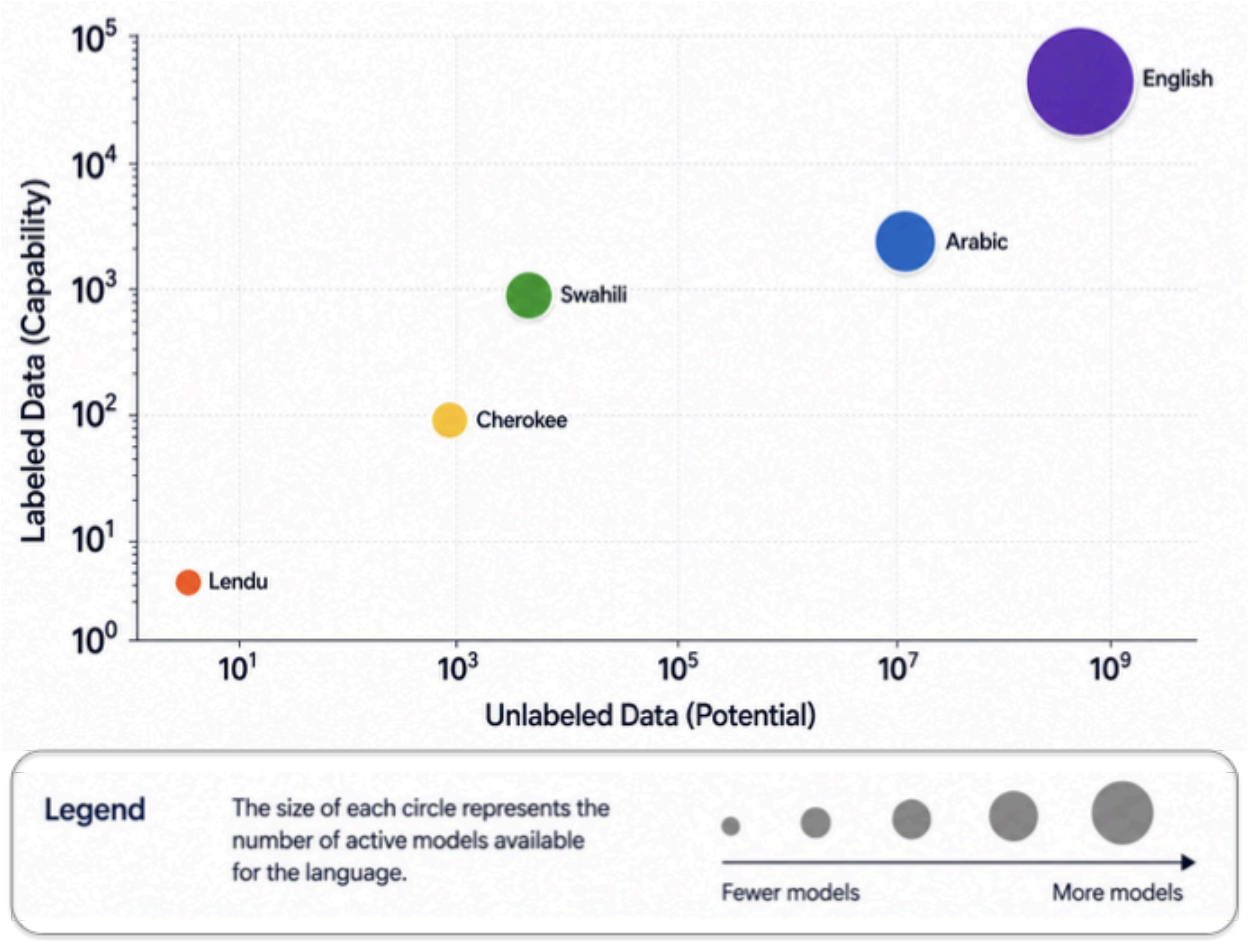
The ARI combines these three variables (L, M, and U) into a single weighted, logarithmic score. Labeled data carries the greatest weight, reflecting its importance in enabling AI functionality.

Because languages vary widely in scale, growth is measured in orders of magnitude rather than raw counts, so smaller and larger languages can be assessed meaningfully on the same scale.



Based on the combined score and per-component minimum floors, a language is placed in one of six classes:

Class	Strategic status
Class 0	Nonexistent. Digitally invisible; home to over a billion people currently excluded.
Class 1	Inert. Trace digital existence; machines essentially cannot "see" the language.
Class 2	Experimental. Brittle performance; requires high human intervention.
Class 3	Emerging. Balanced foundations with established utility.
Class 4	Latent Power. High raw potential; requires specific instructional refining.
Class 5	Saturated. AI is culturally native; resources are self-sustaining.



An ARI Class result tells you four things: what currently exists for the language in AI systems, what is possible today with the data and tools available, what is not yet possible without further investment, and where the specific gaps lie.

A lower class does not mean a language is lacking or needs to change. It reflects its current level of representation and activity in AI systems, no more.

One of the MVD Working Group's outputs is an interactive assessment tool, available at the ARI Assessment Tool. Communities can enter a language to see its current ARI Class and recommended next steps. If the language is not yet in the index, a submission form accepts new entries.

The ARI's full scoring formula, minimum floor thresholds, and per-variable data sources are in Appendix A.

3.2 The AI Capability Model: What AI Functionality the Data Can Support

The ARI describes a language's current state of AI readiness. The AI Capability Model describes what AI functions the language data can support if appropriately structured. The two are related but not interchangeable: a language at a given ARI Class may be capable of supporting one or more tiers in the AI Capability Model, depending on how its data is structured and annotated.

Below we lay out the data required for each tier and what kinds of AI functionality it enables:

- **Tier 1: Basic Discoverability.** Tier 1 establishes a language's initial digital presence. At this level, a modest amount of text (on the order of a few thousand tokens) is segmented into sentences and accompanied by linguistic, contextual, and governance metadata, then published in machine-readable form. The aim is visibility: the language can be indexed by search engines and digital catalogs. Tier 1 does not yet support interactive AI but it makes the language findable.

  Example: a speaker of an underrepresented language searches the web and finds their language listed in dictionary results, Wikipedia articles, and academic publications, where previously the searches returned nothing. New Zealand academic John Moorfield, in collaboration with the Māori Language Commission, created an online Māori-English dictionary and collection of texts, giving the language its first searchable presence online

(see https://maoridictionary.co.nz/). Before this work, Māori text was largely invisible to digital systems.

- **Tier 2: Interactive Tools.** Tier 2 expands the dataset to support interactive applications. It requires larger volumes of text and speech with structured annotations: part-of-speech tags, named entities, dialogue turns, parallel sentence pairs for translation, code-switching indicators, and speech segments with timestamps for automatic speech recognition. Tier 2 supports tools that respond to user input, including basic translation, conversational agents, and speech recognition.

  Example: a community member dictates a message in their first language and a phone transcribes it accurately; a translation app converts a government form from English into the language with reasonable fidelity. The Language Technologies Unit at Bangor University developed a suite of interactive tools for Welsh, including a spelling and grammar checker and a Welsh-English machine translation system. These tools were built using large annotated textual collections with parallel bilingual texts. Welsh speakers can now write, translate, and interact with digital services in their own language.[6]

- **Tier 3: Advanced AI Applications (Generative AI).** Tier 3 enables high-capability AI systems that can produce complex, context-aware, and culturally grounded outputs. Datasets at this level incorporate deep linguistic and semantic representation: semantic role labeling, discourse structure, idiom and metaphor glosses with usage constraints, instruction-tuning pairs, and links to lexical-semantic ontologies. The governance metadata layer deepens correspondingly: machine-readable policy files, explicit consent types and licensing conditions, stewardship roles, and re-consent intervals. Tier 3 supports high-capability AI systems aimed at producing complex, context-aware output.

  Example: a teacher prompts a generative AI to produce a culturally grounded children's story in the language, with regional idioms preserved and characters drawn from local rather than imported cultural references. Using over three billion words of Finnish text, researchers at the University of Turku trained a language model that outperforms general multilingual AI models on Finnish language tasks. Finnish speakers now have access to AI tools that understand and generate texts in their language with greater cultural and linguistic accuracy.[7]

Progression across tiers is not automatic and not required. Communities may choose to remain at a given tier depending on their goals, values, and priorities. Higher tiers do not replace lower ones; they layer on top. Cultural and governance requirements increase alongside technical capability.

---

[6] "Language technology for Welsh as a less-resourced language": https://research.bangor.ac.uk/en/impacts/language-technology-for-welsh-as-a-less-resourced-language-ref202/

[7] Luukkonen et al. (2023), "FinGPT: Large Generative Models for a Small Language", arxiv.org/abs/2311.05640

Full per-tier data requirements (token counts, audio hours, coverage thresholds, quality gates) are in Appendix B and associated other technical requirements are in Appendix E.

3.3 The Cultural and Ethical Gap Analysis

The ARI and AI Capability Model describe what is technically possible. Neither determines if AI output faithfully reflects the nuances of language and culture. A language can have strong technical capacity and produce AI output that is grammatically fluent but culturally flattened.

The Cultural and Ethical Gap Analysis considers readiness at each ARI Class to identify missing or weak cultural and governance elements. It assesses readiness gaps within each technical class rather than producing a separate score. A language may have high technical capacity (for example, ARI Class 4) but significant cultural readiness gaps. The gap analysis reveals what needs to be addressed for ethical AI deployment, and to whose benefit ethical AI deployment is oriented.

CARE-Based Scoring

The analysis uses the CARE Principles. For each, the assessor marks the language as:

- ✓ **Present (Score 4-5)**: the governance practice is robust and functioning.
- ⚠ **Weak (Score 2-3)**: the practice exists but is limited or non-binding.
- ✗ **Gap (Score 0-1):** the practice is missing or token only.

Each principle answers a single question:

- **Collective Benefit (C)**: Does the community receive equitable benefits?
- **Authority to Control (A)**: Does the community have decision-making power over its data?
- **Responsibility (R)**: Are developers accountable to the community?
- **Ethics (E)**: Are cultural protocols and sacred knowledge respected?

Full scoring criteria with example evidence for each principle are in Appendix D.

Cultural Representation Quality

In addition to CARE-based governance assessment, the gap analysis evaluates Cultural Representation Quality: how faithfully AI outputs reflect the language and its culture, using the Cultural Richness scale (See Appendix C for full definitions).

This dimension exists because governance and output quality can come apart. A community may have provided consent and maintained authority over their data while the AI's outputs

misrepresent their culture. CARE scores well; CRC scores poorly. The gap analysis surfaces this gap.

Cultural Representation Quality assessment requires community members with relevant knowledge, including high-context norms and sacred knowledge. It is not a desk-research task. Researchers may provide methodological support; community members lead.

Multi-Stakeholder Approach And Assessment Levels

Different assessors have access to different evidence and bring different perspectives. Community members know informal governance and lived experience but may not have access to external documents. External researchers can access public documents and bring a neutral perspective but may miss informal governance. AI developers know technical implementation and have access to internal documents but may have conflicts of interest. Third-party auditors are independent and can request confidential documents but require resources and community cooperation.

Community members should lead the assessment, with external researchers providing methodological support, not the reverse.

The Cultural and Ethical Gap Analysis can be conducted at three levels of depth:

- **Level 1: Desk Research**. Public documents only (dataset documentation, model cards, academic publications, organization websites, open-source code, Translation Commons Zero to Digital Guidelines). Output: preliminary gap identification.
- **Level 2: Community Self-Assessment.** Adds public-facing documents (annual reports, partnership agreements, funding reports, meeting minutes, newsletters) and confidential records (contracts, data licenses, financial records, internal governance structures). Assessor: community members with governance knowledge. Output: detailed gap analysis with supporting evidence.
- **Level 3: Third-Party Audit.** Adds primary data collection (community interviews, focus groups, surveys, observation of governance processes in practice, data flow audits). Assessor: independent auditor with community cooperation. Output: verified gap analysis with recommendations.

The full assessor-type comparison table, the per-assessor question lists, and the Cultural and Ethical Gap Analysis worksheet template are in Appendix D.

The Cultural and Ethical Gap Analysis requires community members with relevant knowledge, not external researchers working from public documents alone. Class 0 and Class 1 assessments may be desk-based; serious assessment requires community-led evaluation. See §5 for governance modes.

How The Three Elements Relate

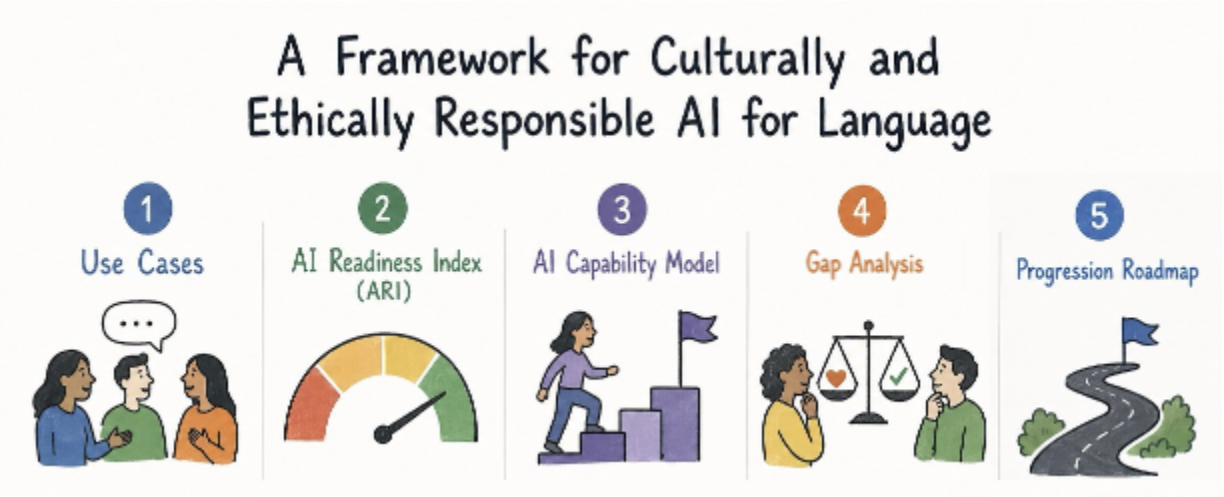
Together, the three elements can help clarify what is needed for a language to be integrated into AI systems. For example:

- A language can be at ARI Class 4 (rich data, many models) and still operate at Tier 1 on the AI Capability Model if the data isn't structured for interactive use.
- A language can be at Capability Tier 3 on the AI Capability Model (technically able to support generative AI) and still produce Cultural Richness Class 2 output if the cultural representation work hasn't been done.
- A language can be culturally well-represented (Cultural Richness Class 4) at modest ARI Class 2 if the small amount of data that exists has been carefully curated by the community.

The path forward (§4) walks a community through the practical decisions: pick a use case, score the ARI Class, pick a target Capability Tier, assess Cultural and Ethical gaps, and build a roadmap.

4. The Path Forward for a Community

The MVD framework walks a community through five steps. The first defines a goal. The next three assess where the language stands across the three lenses introduced in §3. The fifth turns the assessment into a plan.



The pathway is not strictly linear. You can enter at any point if you already have some of the assessments done. You can pause. You can stop at any tier. The framework is designed to support choice, not to compel progression.

Step 1: Define a Use Case

The first step in the pathway is to define how your community wants to use its language in AI-enabled systems, for communication, learning, public life, or storytelling and the arts.

This work should be led by the community itself, often in collaboration with supporting actors such as researchers, institutions, and technical practitioners. The goal is not to define abstract outcomes, but to identify specific ways the language should function in real-world contexts.

A use case is centered on human needs and real-world engagement: who the language is for, what they need to do with it, and in which contexts. At a minimum, each use case should include three elements:

- **Human Purpose.** What meaningful activity does this use case support? (Everyday communication, education, civic access, storytelling, preservation.) For example: a learner asking how to say 'I will see you tomorrow' in Swahili; a small business owner searching for a culturally resonant proverb for a marketing campaign; a community member accessing government services in their first language.
- **Primary Users.** Who is the use case designed for? (Language speakers, learners, educators, community members, institutions, public-facing systems.)
- **Context of Use.** In what real-world setting does this use case occur? (Classroom, home, public services, community spaces, digital platforms.)

	Scenario	What This Enables	Language Complexity
Communication	A person asks a chatbot: *"Can I reschedule my appointment for tomorrow?"* The system understands tone, politeness, and context.	• Asking questions • Clarifying information • Expressing needs	Basic conversational language
Learning	A learner practices dialogue with an AI tutor. "Excuse me, where is the bus station?" The system responds and continues the conversation.	• Practicing real conversations • Learning tone and register • Understanding everyday speech	Structured dialogue and explanation
Civic Participation	A person receives a public health message: *"Free vaccinations available this Saturday at the community clinic."* They can ask questions and understand instructions.	• Access to public services • Understanding announcements • Interacting with instructions	Formal / instructional language
Cultural Expression	A user asks an AI system: *"Tell me a traditional story."* The system generates a culturally meaningful narrative.	• Storytelling • Cultural preservation • Creative expression	Narrative and expressive language

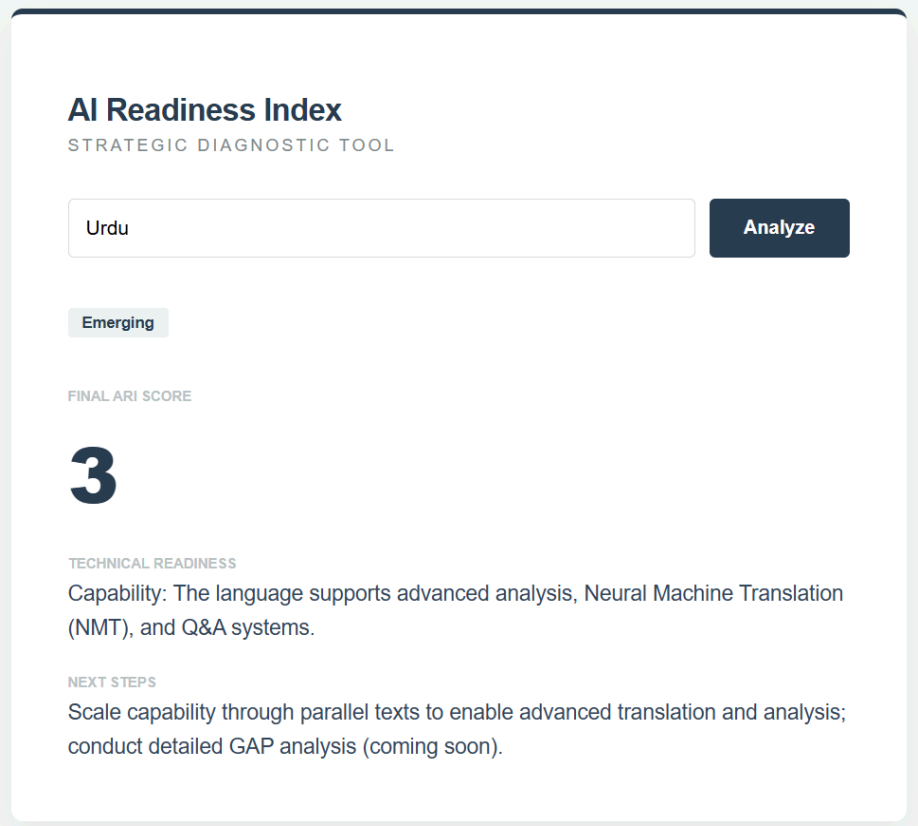
Use cases vary in the level of language technology they require and the amount of data needed to support them. Some applications, such as search and content discovery, can work with relatively small amounts of organized text. Others, such as generative AI, require much larger collections of text and speech data. The AI Capability Model in Section 3.2 connects these data requirements to the types of AI capabilities that can realistically be achieved.

Without clearly defined use cases, language data risks becoming disconnected from real needs: large datasets that aren't built for real communication, AI that generates correct grammar but fails in real-life conversations, speakers who need to switch to English or another dominant language to access digital services, or language that appears in datasets while cultural context disappears. Grounding data work in use cases keeps the framework focused on meaningful participation rather than abstract technical goals.

Step 2: Score the Language's ARI Class

Once the use case is defined, the next step is to understand the language's current position in AI ecosystems. The ARI (§3.1) provides this assessment.

You can score a language's ARI Class using the interactive ARI Assessment Tool. The tool returns the current Class and recommended next steps. Languages not yet in the index can be submitted via the new-language form. The underlying scoring data for all indexed languages is available here. The ARI Tool is considered a working prototype with further iterations as a focus of Phase II work.



Step 3: Pick a Target Capability Tier

The AI Capability Model (§3.2) describes what AI capability the language's data can support. Your target tier follows from your use case: a basic-discoverability use case requires Tier 1; an interactive translation tool requires Tier 2; generative AI requires Tier 3.

If your use case requires a tier above the current ARI Class, Step 5 (the roadmap) is where you plan the progression.

Step 4: Run the Cultural and Ethical Gap Analysis

Reaching a given level of technical capability does not automatically mean a language is ready for AI deployment. A system may be technically functional while still misrepresenting cultural meaning, violating community expectations, or operating without appropriate consent or oversight.

The Cultural and Ethical Gap Analysis assesses whether the cultural, ethical, and governance conditions required for responsible AI use are in place. While the ARI Class evaluates technical presence and capability, the gap analysis evaluates whether progression is appropriate, responsible, and aligned with community values.

The analysis uses the CARE Principles as its foundation:

- **Collective Benefit.** Are benefits shared equitably with the community?
- **Authority to Control.** Does the community have decision-making power over its data?
- **Responsibility.** Are there mechanisms for accountability, review, and redress?
- **Ethics**. Are cultural norms, protocols, and sensitivities respected?

Each dimension is assessed for whether the practices around the dataset are present and functioning, partially established but weak, or missing entirely. This allows communities and supporting actors to identify not just whether governance exists, but whether it is sufficient for the intended level of AI capability.

The Cultural and Ethical Gap Analysis also evaluates Cultural Representation Quality using the Cultural Richness Class scale (§3.3). A system may meet technical requirements but still flatten cultural meaning, misinterpret context or tone, or fail to capture idioms, metaphors, or social norms.

The outcome is a structured identification of two types of gaps:

- **Governance gaps:** lack of consent frameworks, absence of community control, unclear ownership or licensing, missing accountability mechanisms.
- **Cultural gaps:** insufficient contextual data, misrepresentation of meaning, lack of culturally grounded annotation or review.

These gaps indicate what you must address before progressing to higher levels of AI capability. The full assessment process, evidence-gathering levels, and a worksheet template are in §5 and Appendix D.

Step 5: Build a Roadmap

Once you have defined a use case, scored the ARI Class, chosen the target Capability Tier, and run the Cultural and Ethical Gap Analysis, the final step is to turn the assessment into a roadmap[8].

A progression roadmap is a structured plan that identifies what is missing, defines what actions are required, and sequences those actions over time. It serves as a bridge between assessment and implementation, ensuring that development is purposeful, coordinated, and aligned with community priorities.

A complete roadmap typically includes a gap summary (from Step 4), priority actions tied to the chosen use case and capability target, a sequencing plan, role assignments per the ecosystem described in §2, and milestones at which progress is reviewed and the path forward is confirmed, paused, or adjusted.

A Worked Example: Swahili

The framework was applied to Swahili as a worked example.

ARI Class: 3 (Emerging: balanced foundations with established utility).

Cultural and Ethical Gap Analysis:

CARE Principle	Status	Notes
Collective Benefit	Present	Community-led organizations receive funding. Training programs for community members. Educational resources developed.
Authority to Control	Present	Community representatives on the development team. Consent protocols documented. Community input on data-use decisions.
Responsibility	Weak	Annual reporting exists but no formal contestability process. Accountability structures need strengthening.

[8] The MVD framework provides direction by calling for a roadmap informed by the ARI Class, Target Capability Tier, and Cultural and Ethical Governance Gap Analysis. However, as noted in section 6, the final roadmap structure and online interface have not been created during this phase of work.

CARE Principle	Status	Notes
Ethics	Present	Cultural protocols documented and followed. Community elders consulted. Respectful approach to language preservation.
Cultural Representation Quality	Current Class 3 / Expected Class 3	Cultural Contextualisation is broadly achieved: the AI handles regional knowledge and common-sense cultural references accurately. Figurative and Pragmatic Fluency (Class 4) has not yet been assessed by community members with specialist knowledge of Swahili idiom and high-context communication. Community-led assessment required before claims of Class 4 readiness.

Summary: Swahili shows strong technical capacity at ARI Class 3 and broadly sound governance, with a minor accountability gap. On Cultural Representation Quality, Swahili meets the expected Class 3 standard, but progression to Class 4 requires community-led assessment. Swahili is conditionally ready for ethical deployment; community self-assessment of cultural richness should follow.

Non-Linear Progression And The Gated Approach

Progression is not always linear. Governance gaps may need to be addressed before technical advancement. Technical work may begin while governance structures are still developing. Communities may choose to pause or remain at a given level. This flexibility is intentional: development is guided by choice and readiness, not by external pressure to advance.

The MVD adopts a gated approach to capability progression. Advancement is not determined by technical capacity alone. It depends on whether the necessary data and systems are available (the ARI Class), AND whether the required cultural and governance conditions are in place (the Cultural and Ethical Gap Analysis). Where significant gaps exist, particularly in governance, progression is limited or deferred until those gaps are addressed. This ensures that AI development does not proceed in ways that are extractive, misrepresentative, or misaligned with community priorities.

5. Governance and Consent

Governance is not an add-on to the MVD framework. It is foundational. Without it, efforts to improve AI readiness for a language can produce extractive data practices, loss of cultural meaning, and systems that fail the communities whose languages they rely on. With it,

communities can shape how their language is represented, used, and sustained in digital environments.

This section sets out the governance principles the MVD aligns with, the participation modes that make those principles operational, and the Cultural and Ethical Gap Analysis methodology used to assess whether the conditions for ethical AI deployment are actually in place.

5.1 Principles for Language Data Governance

The MVD framework aligns with established principles of Indigenous and community data governance:

- CARE (Collective Benefit, Authority to Control, Responsibility, Ethics), Global Indigenous Data Alliance.
- OCAP® (Ownership, Control, Access, Possession), First Nations Information Governance Centre.
- UNDRIP (United Nations Declaration on the Rights of Indigenous Peoples), particularly the articles on cultural heritage and language.
- Maiam nayri Wingara Indigenous Data Sovereignty Collective principles, Aboriginal and Torres Strait Islander data governance.
- FPIC (Free, Prior, and Informed Consent) as the consent norm for data collection from Indigenous and community contexts.

These principles provide a foundation for ensuring that language data is managed in ways that respect community rights and interests. Governance frameworks should be adapted to local contexts and may combine formal legal structures with community-based norms and practices.

The MVD framework is intended to align with emerging global AI governance and policy frameworks, including the UNESCO Recommendation on the Ethics of Artificial Intelligence and broader international standards on digital inclusion and Indigenous data rights. Periodic review should consider these evolving policy environments to ensure alignment and enable communities to draw on existing protections and governance mechanisms.

Governance is not a burden communities carry alone. The actors who design, fund, and deploy AI systems share responsibility for ensuring the conditions of ethical use are met. Communities are partners and decision-makers.

Consent and compensation are inseparable. Meaningful consent requires that communities understand how their data will be used and how they will benefit; compensation without prior, informed consent is extraction rather than partnership.

Governance considerations apply at every stage of the MVD pathway (§4): use case definition, ARI scoring, target Capability Tier selection, Cultural and Ethical Gap Analysis, and roadmap planning. Governance is not a discrete step; it is the lens through which every step is taken.

5.2 Participation and Co-Design

Governance is operationalized through participation. Communities must have meaningful opportunities to shape decisions about how their language is represented and used in AI systems. This includes:

- Participatory data governance structures that give communities oversight of data collection and use.
- Co-design processes where community members collaborate in defining datasets, tools, and use cases.
- Feedback loops and review mechanisms that keep systems aligned with community needs over time.

Engagement approaches should be adapted to local contexts. The ability to participate in governance decisions should not be conditional. In many cases, this means:

- Oral consultations rather than text-based surveys.
- Multilingual facilitation.
- Low-bandwidth or offline participation formats.
- Culturally appropriate consent practices.

Participation must be sustained across the full lifecycle of language data and AI systems, not limited to initial consultation.

5.3 Trade-Offs and Power Dynamics

Language data governance often involves navigating trade-offs and power imbalances. Examples include external organizations seeking access to language data for model development, large technology platforms shaping how languages are represented online, and resource constraints limiting community participation.

In these contexts, governance frameworks must ensure that communities retain authority over their language data, benefits are equitably distributed, and risks of misuse or misrepresentation are mitigated. This may require negotiation, capacity-building, and the development of new institutional arrangements.

5.4 Outcomes and Ongoing Review

When governance is strong, AI readiness efforts are more likely to result in:

- Language technologies that reflect real-world usage and cultural context.
- Increased trust and participation from language communities.
- Sustainable data ecosystems that can evolve over time.
- Equitable distribution of benefits derived from language data.

Conversely, weak governance can lead to systems that are technically functional but socially misaligned or harmful.

Governance also includes planning for end-of-life. When a project ends, a developer withdraws, or funding stops, the data and stewardship arrangements need explicit exit conditions. Datasets should record stewardship transfer plans (who retains the data, who holds decision authority, and under what license terms it persists or is withdrawn), agreed at project inception rather than improvised at close. This connects directly to the CARE Responsibility principle: stewardship is not optional, and gaps in archival accountability are themselves governance gaps.

Governance and ecosystem development are ongoing processes. The Working Group recommends periodic review of governance structures to ensure the framework remains responsive to the communities it serves rather than the systems it governs.

6. Looking Ahead: Phase 2 and Next Steps

The MVD framework set out in this report is the output of Phase 1. Phase 2 will turn the framework from a shared reference into operational use. These are presented as options, not as a planned sequence; any one could anchor Phase 2 on its own.

1. **Share Phase 1 artifacts with the CODI community for feedback and iteration.** The Phase 1 artifacts (this report, the AI Readiness Index Assessment Tool and the Cultural and Ethical Gap Analysis) are starting points, not finished products. One direction is to share them with the broader CODI audience and incorporate feedback from communities, researchers, and practitioners who put them into use, with successive versions shaped by that input.

2. **Consolidate forkable open-source infrastructure.** A similar direction would be to consolidate the technical artifacts into a single GitHub repository that combines the theoretical framework with a base project communities can fork and adapt. The repository would host the Tier 1–3 JSON schemas, the CODI Cultural-Pragmatic and Governance Ontologies, validation tooling, and reference samples, so that organizations applying the MVD start from a working baseline rather than rebuilding from scratch. A scaffold already exists at the CODI-MVD repository.

3. **Operationalize and implement the framework for one case study.** Select a candidate language community and assemble a complete ecosystem of actors around it, drawing from within the CODI network. Community representatives, governance experts, researchers, and technical partners can each contribute a piece of the implementation, leveraging the work they are already doing in low-resource language and Indigenous data sovereignty. A pilot would walk the full pathway: define use cases with the community, score the language with the ARI Assessment Tool, target a Capability Tier, run the Cultural and Ethical Gap Analysis, and follow the resulting roadmap end to end. The outcome is a concrete worked example and a partnership template for replicating the model with additional language communities. The Operational Model for Low-Resource Language AI offers a possible concrete path.

4. **Consolidate the framework into a unified digital platform**. Evolve the individual conceptual components into a single, interactive online platform for stakeholders. This platform would house the entire workflow in one location. Specifically, this phase must finalize the roadmap structure itself and explicitly map how these moving parts connect. Stakeholders would then be guided through each phase dynamically, using built-in digital tools to input their data, assess their language's current state, and automatically generate a tailored, actionable roadmap for dataset creation, cultural integration, and ethical governance on a single interface.

Which of these directions Phase 2 prioritizes, and in what sequence, is a question of future work. Communities, researchers, technologists, and funders interested in shaping that decision are invited to share feedback on this report, contribute to any open repository when it launches, and join the CODI newsletter for ongoing updates.

References

Apple Support. “How to Get Apple Intelligence.” *Apple Support*, 2026, https://support.apple.com/en-us/121115.

Bender, E. M. “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.” *Transactions of the Association for Computational Linguistics*, 2018, https://doi.org/10.1162/tacl_a_00041.

Bird, Steven. “Decolonising Speech and Language Technology.” *Information Processing & Management*, 2020, https://www.sciencedirect.com/science/article/pii/S0306457320307488.

Boersma, Paul, and David Weenink. *Praat: Doing Phonetics by Computer*. https://www.fon.hum.uva.nl/praat/.

CARE Principles for Indigenous Data Governance. *Global Indigenous Data Alliance (GIDA)*, 2018, https://www.gida-global.org/careprinciples.

DAMA International. *The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK2)*. Technics Publications, 2017, https://dama.org/dmbok2r-infographics/

Data Nutrition Project. “The Dataset Nutrition Label.” *Data Nutrition Project*, 2026, https://datanutrition.org/.

Ethnologue. *Languages of the World*. 2026, https://www.ethnologue.com/.

First Nations Information Governance Centre (FNIGC). “The First Nations Principles of OCAP®.” *FNIGC*, 1998, https://fnigc.ca/ocap-training/.

Gebru, Timnit, et al. “Datasheets for Datasets.” *Communications of the ACM*, 2021, https://arxiv.org/abs/1803.09010. Originally published 2018.

GIDA. “CARE Principles for Indigenous Data Governance.” *Global Indigenous Data Alliance*, 2018, https://www.gida-global.org/careprinciples.

Global Indigenous Data Alliance (GIDA). *Global Indigenous Data Alliance*. 2019, https://www.gida-global.org/.

Google. “Africa's Digital Decade: AI Upskilling and Expanding Speech Technology.” *Google Blog*, 2024, https://blog.google/company-news/inside-google/around-the-globe/google-africa/africas-digital-decade/.

The Guardian. “Group of High-Profile Authors Sue Microsoft over Use of Their Books in AI Training.” *The Guardian*, 2025, https://www.theguardian.com/technology/2025/jun/26/microsoft-ai-authors-lawsuit.

Hershcovich, Daniel, et al. “Challenges and Strategies in Cross-Cultural NLP.” *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, https://aclanthology.org/2022.acl-long.482.pdf.

Hovy, D., and S. Prabhumoye. “Five Sources of Bias in Natural Language Processing.” *Language and Linguistics Compass*, 2021, https://doi.org/10.1002/lnc3.12432.

Hutchinson, Ben, et al. “Towards Accountability for Machine Learning Datasets.” *Proceedings of FAccT 2021*, 2020, https://arxiv.org/abs/2010.13561.

Indiana University at Indianapolis. “AI-AI Esthetic Collaboration with Explicit Semiotic Awareness and Emergent Grammar Development.” 2025, https://arxiv.org/abs/2508.20195.

Internet Engineering Task Force. *RFC 5646 (BCP 47): Tags for Identifying Languages*. https://www.rfc-editor.org/info/bcp47.

Joshi, Pratik, et al. “The State and Fate of Linguistic Diversity and Inclusion in the NLP World.” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, https://aclanthology.org/2020.acl-main.560/.

JSON Schema. *JSON Schema Specification, Draft-07*. https://json-schema.org/specification-links.html#draft-7.

Jurafsky, D., and J. H. Martin. *Speech and Language Processing (3rd ed. draft)*. Stanford University, 2025, https://web.stanford.edu/~jurafsky/slp3/.

LAMARR Institute. “Ethical Use of Training Data: Ensuring Fairness and Data Protection in AI.” *LAMARR Institute*, 2024, https://lamarr-institute.org/blog/ai-training-data-bias/.

Liu, C. C., et al. “Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art.” *Transactions of the Association for Computational Linguistics*, 2025, https://aclanthology.org/2025.tacl-1.31.pdf.

Maiam nayri Wingara Indigenous Data Sovereignty Collective. *Indigenous Data Governance Framework*. 2018, https://www.maiamnayriwingara.org/.

Masakhane. “Community-Driven NLP for African Languages.” 2020, https://arxiv.org/abs/2003.11529.

Masakhane Research Foundation. *Masakhane Research Foundation*. Ongoing, https://www.masakhane.io/.

Max Planck Institute for Psycholinguistics. *ELAN: Linguistic Annotator*. https://archive.mpi.nl/tla/elan.

Mendonça, F., et al. “Data Cooperatives: Democratic Models for Ethical Data Stewardship.” 2025, https://arxiv.org/abs/2504.10058.

Mitchell, T. M. *Machine Learning*. McGraw-Hill, 1997.

National Indigenous Australians Agency (NIAA). *Framework for Governance of Indigenous Data*. 2024, https://www.niaa.gov.au/sites/default/files/documents/2024-05/framework-governance-indigenous-data.pdf.

The National. “Should We Let Foreign Linguists into PNG?” *The National*, 2026, https://www.thenational.com.pg/should-we-let-foreign-linguists-into-png/.

National Statistical Office of Papua New Guinea. *Papua New Guinea National Census*. 2024, https://www.nso.gov.pg/statistics/population/.

NLLB Team. “No Language Left Behind: Scaling Human-Centered Machine Translation.” *Meta AI Research*, 2022, https://research.facebook.com/publications/no-language-left-behind/.

Nore, S. “A Paradigm Shift: Political Outlook on Digital Transformation and Internet Development Challenges in the Solomon Islands.” 2025, https://netmission.asia/2025/02/19/a-paradigm-shift-political-outlook-on-digital-transformation-and-internet-development-challenges-in-the-solomon-islands-songo-nore/.

OECD. “AI Capability Indicators.” *OECD*, 2025, https://www.oecd.org/en/publications/introducing-the-oecd-ai-capability-indicators_be745f04-en/full-report/component-5.html.

OECD. “AI Principles.” *OECD*, 2024, https://www.oecd.org/going-digital/ai/principles/.

Open Knowledge Foundation. “Open Definition 2.1.” *Open Knowledge Foundation*, 2015, https://opendefinition.org/.

Paullada, A., I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. “Data and Its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research.” *Patterns*, 2021, https://doi.org/10.1016/j.patter.2021.100336.

schema.org. “Dataset.” *Schema.org*, https://schema.org/Dataset.

SIL International. *ISO 639-3: Codes for the Representation of Names of Languages*. https://iso639-3.sil.org/.

Sitaram, S., K. R. Chandu, S. K. Rallabandi, and A. W. Black. “A Survey of Code-Switched Speech and Language Processing.” *arXiv*, 2019, https://arxiv.org/abs/1904.00784.

SPDX. *SPDX License List*. https://spdx.org/licenses/.

Text Encoding Initiative Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. https://tei-c.org/release/doc/tei-p5-doc/en/html/.

Translation Commons. “Zero to Digital Resources.” *Translation Commons*, 2026, https://translationcommons.org/zero-to-digital-resources.

UNESCO. “Global Roadmap on Multilingualism in the Digital Era: Advancing the Role of Language Technologies.” *UNESCO*, 2025, https://www.unesco.org/en/articles/unesco-launches-global-roadmap-multilingualism-digital-era.

UNESCO. “Languages Matter: Global Guidance on Multilingual Education.” *UNESCO*, 2025, https://unesdoc.unesco.org/ark:/48223/pf0000392477.

UNESCO. “Recommendation on the Ethics of Artificial Intelligence.” *UNESCO*, 2021, https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence.

UNESCO Ad Hoc Expert Group on Endangered Languages. *Language Vitality and Endangerment*. 2003, https://ich.unesco.org/doc/src/00120-EN.pdf.

United Nations. *United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP).* 2007, https://www.un.org/development/desa/indigenouspeoples/declaration-on-the-rights-of-the-indigenous-peoples.html.

United Nations. *United Nations Declaration on the Rights of Indigenous Peoples.* United Nations, 2007, https://www.un.org/development/desa/indigenouspeoples/declaration-on-the-rights-of-the-indigenous-peoples.html.

Universal Dependencies. *CoNLL-U Format*. https://universaldependencies.org/format.html.

Universal Dependencies Project. “Universal POS Tags.” *Universal Dependencies*, https://universaldependencies.org/u/pos/.

W3C. *Data Catalog Vocabulary (DCAT) Version 3*. 2024, https://www.w3.org/TR/vocab-dcat-3/.

W3C Ontology-Lexicon Community Group. *OntoLex-Lemon: Lexicon Model for Ontologies*. 2016, https://www.w3.org/2016/05/ontolex/.

Wanjawa, B., L. Wanzare, F. Indede, O. McOnyango, E. Ombui, and L. Muchemi. “Kencorpus: A Kenyan Language Corpus of Swahili, Dholuo and Luhya for Natural Language Processing Tasks.” *arXiv*, 2022, https://arxiv.org/abs/2208.12081.

XDA Developers. “Here's Everything Google Added to Gemini in January 2026.” *XDA Developers*, 2026, https://www.xda-developers.com/everything-google-added-to-gemini-january-2026/.

Yip, M. *Tone*. Cambridge University Press, 2002.

Zamir, S. W., et al. “BYOL: Bring Your Own Language Into LLMs.” 2026, https://arxiv.org/abs/2601.10804.

Glossary

AI (Artificial Intelligence)

The capability of computational systems to perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception, and decision-making. It is a multidisciplinary field spanning computer science, mathematics, and engineering.

Annotation

The process of labeling, tagging, or adding structure and meaning to raw data (e.g., identifying parts of speech or marking cultural context) so it can be used for analysis or training machine learning models.

ARI (AI Readiness Index)

A weighted logarithmic model that measures a language's current presence in AI ecosystems. The ARI assesses three components: Labeled Data (L), Model Activity (M), and Unlabeled Data (U). It places a language in one of six ARI Classes (Class 0 through Class 5). See §3.1 and Appendix A.

ARI Class

The output of the ARI: a placement in Class 0 (Digitally Invisible) through Class 5 (Saturated) based on the L / M / U weighted score and the per-component Minimum Floors.

ASR (Automatic Speech Recognition)

Technology that converts spoken language (audio) into written text using artificial intelligence models.

Benchmark

A set of defined indicators or criteria used to evaluate whether a dataset meets specific standards or levels of readiness for digital or AI applications.

Bias

Systematic errors or imbalances in data or algorithms that can lead to unfair, inaccurate, or discriminatory outcomes.

BYOL (Bring Your Own Language Into LLMs)

An approach that enables communities to integrate and adapt their own languages into large language models by contributing data, tools, and linguistic resources.

Capability Tier

A three-level scale (Tier 1 Basic Discoverability, Tier 2 Interactive Tools, Tier 3 Generative AI) describing what AI applications a language's dataset can support given its annotation depth, governance metadata, and structural readiness. See §3.2 and Appendix B.

Cultural and Ethical Gap Analysis

The MVD assessment method that evaluates cultural and governance readiness alongside technical readiness. Uses the CARE Principles for governance scoring and the Cultural Richness Class for output-quality scoring. See §3.3 The Cultural and Ethical Gap Analysis.

CARE Principles (Collective Benefit, Authority to Control, Responsibility, Ethics)

A framework for Indigenous data governance that emphasizes collective rights, community control, ethical use, and responsibility in data practices. Published by the Global Indigenous Data Alliance (GIDA).

CODI (Coalition On Digital Impact)

A non-profit organization focused on advancing inclusive digital ecosystems, particularly for underrepresented languages and communities.

CODI-CPO (CODI Cultural-Pragmatic Ontology)

A JSON ontology, published in the CODI-MVD GitHub repository, that defines high-level cultural categories such as Sacred Knowledge, Social Roles, and Knowledge Types.

CODI-GOV (CODI Governance Ontology)

A JSON ontology, published in the CODI-MVD GitHub repository, that defines authority over data and specific usage rights.

CoNLL-U

A tab-separated text format used by Universal Dependencies for exchanging linguistically annotated corpora.

Critical Languages

Languages identified as essential for development, governance, or service delivery in a given region, particularly where the absence of digital tools impedes access to public services or civic participation.

Cultural Richness Class

A six-level scale (Class 0 Digital Invisibility through Class 5 Representational Fidelity) describing how faithfully AI outputs reflect a language and its culture. See §3.3 and Appendix C.

Cultural Sensitivity

The practice of ensuring that data collection, use, and technology design respect cultural norms, traditions, values, and community expectations.

Data Governance

The framework of policies, processes, and standards used to manage data responsibly, including issues of ownership, access, security, and compliance.

Data Stewardship

The responsible management, oversight, and protection of data throughout its lifecycle, ensuring its quality, integrity, and appropriate use.

Dataset

An organized collection of data used for analysis, digital applications, or training artificial intelligence systems.

DCAT (Data Catalog Vocabulary)

A W3C Recommendation for describing datasets and data services to enable interoperable catalog publication across the web. The MVD framework uses DCAT alongside schema.org/Dataset for federated discovery.

Digital Inclusion

Efforts to ensure equitable access to digital technologies, resources, and opportunities for all communities, regardless of language, geography, or socioeconomic status.

ELAN

A free annotation tool for speech and video data, developed by the Max Planck Institute for Psycholinguistics. The MVD framework uses ELAN's .eaf format for time-aligned speech transcription.

Ethical Framework

A structured set of principles and guidelines that govern responsible data collection, sharing, **and use, including privacy, consent, cultural sensitivity, and community ownership.**

F1 score

A measure of accuracy combining precision and recall as their harmonic mean. Reported on a 0–1 scale; higher values indicate better performance on tasks such as part-of-speech tagging and named entity recognition.

First Nations

A term used to identify Indigenous peoples in Canada who are neither Inuit nor Métis.

FPIC (Free, Prior, and Informed Consent)

A principle ensuring that communities give informed and voluntary consent before their knowledge, data, or resources are collected, used, or shared. Codified in UNDRIP.

Generative AI

A class of artificial intelligence systems that can create new content (e.g., text, audio, images) based on patterns learned from existing data.

High-Resource Languages

Languages with extensive digital linguistic resources, including large text corpora, speech data, annotated datasets, and well-developed AI tools. English, Mandarin, Spanish, and French are commonly described as high-resource.

IAA (Inter-Annotator Agreement)

A measure of how consistently independent annotators apply the same labels to the same data. Often reported using Cohen's kappa or similar metrics, with values closer to 1.0 indicating greater consistency.

Interoperability

The ability of different systems, platforms, or datasets to work together effectively through shared standards, formats, and protocols.

ISO 639-3

An international standard for three-letter codes that uniquely identify individual languages. The MVD framework uses ISO 639-3 as the base language identifier, complemented by BCP-47 for dialect, script, and region subtags.

JSON Schema Draft-07

A specification for validating the structure of JSON documents. The MVD framework uses JSON Schema to deterministically check metadata structure and governance fields before data is accepted into the stable library.

Kappa (Cohen's kappa)

A statistical measure of agreement between annotators that adjusts for chance agreement. Commonly used to evaluate inter-annotator agreement in linguistic annotation.

Language Corpus

A large, structured collection of spoken or written language used for linguistic research or training language technologies.

Language Vitality and Sovereignty Diagnostic

A voluntary secondary diagnostic for ARI Class 0 and Class 1 languages, used to surface "Hidden Data" (archival, oral, or private records) that public aggregators do not index.

LLMs (Large Language Models)

Advanced AI models trained on large volumes of text data to understand and generate human language.

Machine Learning

A method of artificial intelligence in which systems learn patterns from data to perform tasks without being explicitly programmed.

Maiam nayri Wingara

An Indigenous Data Sovereignty Collective that publishes principles for Aboriginal and Torres Strait Islander data governance, complementary to CARE and OCAP.

Metadata

Structured information that describes and provides context about data (e.g., language, source, authorship, date), enabling organization, discovery, and management.

MVD (Minimum Viable Dataset)

A framework for linguistic and cultural empowerment that aligns minimum language dataset requirements with progressive stages of readiness for digital and AI applications.

MVD WG (Minimum Viable Dataset Working Group)

A group of selected experts representing a wide range of experience in cultural and linguistic representation, AI and data governance, ethics, digital rights, and policy, assembled by the Coalition On Digital Impact (CODI) to define what a Minimum Viable Dataset for equitable

cultural and linguistic representation and participation in today's digital and AI-driven world looks like.

NER (Named Entity Recognition)

An NLP task that identifies and classifies named entities in text, such as people, places, organizations, and dates.

NLP (Natural Language Processing)

A field of artificial intelligence focused on enabling computers to understand, interpret, and generate human language.

OCAP® Principles

Ownership, Control, Access, and Possession. A set of First Nations data governance principles that define how data should be collected, protected, used, and shared. Published by the First Nations Information Governance Centre (FNIGC).

OntoLex-Lemon

A W3C Community Report model for representing lexicons as linked data, with modules for variation and translation that are particularly useful in multilingual and low-resource contexts.

Open Data

Data that is freely available for anyone to access, use, modify, and share, typically with minimal restrictions.

Open Standards

Publicly available specifications and protocols that enable systems and data to be interoperable and consistently implemented.

Oral-Tradition Languages

Languages in which knowledge, history, and cultural practices are primarily transmitted through spoken communication rather than written forms.

POS (Part-of-Speech) Tagging

An NLP task that assigns a grammatical category (noun, verb, adjective, etc.) to each token in a text.

Praat

A free phonetic analysis tool that supports spectrogram inspection, pitch and formant tracking, and multi-tier labeling of speech recordings. The MVD framework uses Praat's .TextGrid format for time-aligned transcription.

Provenance

The documented history of a dataset, including its origin, ownership, and any changes or reviews over time.

Responsive Tools

Interactive digital technologies that respond to user input, such as voice assistants, translation systems, accessibility tools, and conversational AI.

Schema

A structured set of rules that defines how data is organized, formatted, and represented in a dataset or file.

schema.org/Dataset

A widely used vocabulary for describing datasets in machine-readable form. Web search engines and data portals consume schema.org/Dataset metadata to index and surface datasets.

Search & Discovery

The ability for languages and their content to be indexed, found, and accessed through search engines and digital platforms.

SPDX (Software Package Data Exchange) Identifiers

A standardized list of license identifiers (e.g., CC-BY-NC-4.0, MIT) used to label license terms in machine-readable form.

SSRL (Semantic Role Labeling)

An NLP task that identifies the predicate-argument structure of a sentence: who did what to whom, where, and when.

TEI (Text Encoding Initiative)

A standard for encoding texts in digital form using structured markup, widely used in digital humanities and linguistic research. The MVD framework uses TEI P5 for long-form or edition-level textual data.

Token

A basic unit of language used in AI systems, roughly corresponding to a word or part of a word.

Tonal Language

A language in which differences in pitch or tone change the meaning of words.

Training Data

Data used to train machine learning models to recognize patterns and make predictions.

UD (Universal Dependencies)

A framework for consistent grammatical annotation across languages, including parts of speech and syntactic relationships. The MVD framework adopts UD's 17 Universal POS tag categories and its morphological feature inventory.

UD UPOS (Universal Part-of-Speech Tags)

A standardized set of core grammatical categories used in Universal Dependencies annotation.

UNDRIP (United Nations Declaration on the Rights of Indigenous Peoples)

A UN declaration adopted in 2007 that affirms Indigenous peoples' rights, including rights to cultural heritage, languages, and traditional knowledge.

Use Cases

Specific scenarios or applications that describe how a dataset, tool, or system is intended to be used.

Velocity Multiplier (V = 1.15)

A bonus applied within the ARI score when a language's model activity (M) exceeds twice the available labeled data (L), recognizing communities over-performing on model development relative to raw data. See Appendix A.3.3.

WER (Word Error Rate)

A standard accuracy metric for automatic speech recognition. Calculated as the number of word-level substitutions, insertions, and deletions divided by the total number of reference words; lower values indicate better performance.

Appendices

Appendix A. The Ai Readiness Index (ARI), Full Scoring

A.1 Overview

The AI Readiness Index (ARI) is a weighted logarithmic model designed to measure the "Operational Health" of a language within modern Artificial Intelligence ecosystems. The framework builds on the foundational Language Taxonomy introduced by Joshi et al. (2020) and categorizes a language based on its Ecosystem Vitality across three dimensions: Potential (unlabeled data), Capability (labeled data), and Activity (model count). In addition to the labeled and unlabeled data captured by Joshi et al., the ARI accounts for model activity and introduces a weighted system that distinguishes raw data availability from active language-model development.

A.2 ARI Classes and Minimum Floors

The final ARI Score determines a language's placement. Moving from one Class to the next reflects a roughly ten-fold increase in resource availability (an order of magnitude). To advance to a higher Class, a language must meet the ARI Score Range and clear the Minimum Floor for all three mathematical variables (L, M, and U).

Classification	Strategic Status	L-Floor (Labeled Data)	U-Floor (Unlabeled Data)	M-Floor (Model Count)	ARI Score Range
Class 0	Nonexistent: digitally invisible; home to a substantial share of the world's population currently excluded from AI participation.	0	0	0	< 0.1
Class 1	Inert: trace digital existence; machines essentially cannot "see" the language.	1	1	1	0.1–1.0

Classification	Strategic Status	L-Floor (Labeled Data)	U-Floor (Unlabeled Data)	M-Floor (Model Count)	ARI Score Range
Class 2	Experimental: brittle performance; requires high human intervention.	10	100	10	1.0–2.5
Class 3	Emerging: balanced foundations with established utility.	100	100,000	100	2.5–3.5
Class 4	Latent Power: high raw potential; requires specific instructional refining.	500	1,000,000	1,000	3.5–4.5
Class 5	Saturated: AI is culturally native; resources are self-sustaining.	1,000	10,000,000	5,000	> 4.5

A.3 ARI Mathematical Architecture

The index is calculated using a weighted logarithmic formula:

$$ARI = (0.50 \cdot \log_{10}(L) + 0.30 \cdot \log_{10}(M) + 0.20 \cdot \log_{10}(U)) \cdot V$$

Output is mapped to the 0–5 Class via the Score Range table above.

When any of L, M, or U is zero, the corresponding logarithm is undefined and ARI is defined to be zero. In practice this only occurs at Class 0, where the Bottleneck Rule (Minimum Floor of L ≥ 1, M ≥ 1, U ≥ 1 to reach Class 1) keeps the language at Class 0 regardless of the formula's value, so the edge case does not affect placement for any other Class.

A.3.1 The Logarithmic Scale (log_{10})

A base-10 logarithm accounts for the massive scale of linguistic data and the diminishing returns of growth. Progress is measured in orders of magnitude: the jump from 10 to 100 datasets is treated as equally transformative as the jump from 1,000 to 10,000. This allows languages of massive scale (e.g. 10^9 documents) to be compared to smaller languages (e.g. 10^2 documents) on a single 0–5 point scale.

A.3.2 Weighted Variables

The following variables are weighted by their necessity for creating a functional large language model:

- L (Labeled Data / Capability), 50%. Represents Instructions and is the primary driver of AI utility. Source: Linguistic Data Consortium (LDC) and Hugging Face Hub.
- M (Model Activity / Implementation), 30%. Represents Proof of Concept; identifies how much human effort is being spent to build tools for the language (a "Kinetic Filter"). Source: Hugging Face Hub.
- U (Unlabeled Data / Potential), 20%. Represents Raw Potential and provides the structural foundation. Source: OSCAR (Open Super-large Crawled Aggregated coRpus) web-scale crawls.

A.3.3 The Velocity Multiplier (V = 1.15)

A 1.15× bonus is applied to a "High-Velocity" community when the number of public models exceeds twice the available labeled datasets, i.e. M > (L × 2). This acts as a Kinetic Filter to recognize communities that are over-performing relative to their raw data.

A.3.4 Placement and Indexing Logic: the Bottleneck Rule

To ensure technical validity, a language must clear the Minimum Floors (L-Floor, M-Floor, U-Floor) to advance to a Class. If a language has a high ARI score for a higher Class (e.g. Class 4) but fails to meet the Minimum Floor for any single variable, it is down-rounded to the highest Class where it meets all requirements. Languages that clear the score range for a higher Class but are held back by a single floor are tagged "Emergent," identifying them as high-priority candidates for targeted data investment.

A.4 The Voluntary Step 2 Diagnostic (Readiness Index)

For languages categorized as Class 0 or Class 1 (indicating a "Public Data Gap"), the public ARI score is a readiness indicator, not a final judgment. A Secondary Diagnostic is available and entirely voluntary, to establish a language's true Readiness Index. This audit seeks to uncover "Hidden Data" including archival, oral, or private records that public aggregators often fail to index.

The Secondary Diagnostic is currently realized through the Language Vitality and Sovereignty Diagnostic Google Form.

A.5 Scope and Limitations

The AI Readiness Index is intended as a recommended starting point to inform strategic decision-making and should not be used as a definitive determination of a language's long-term viability or value. The score is one input among many for community planning.

Appendix B. Capability Tier Data Requirements

B.1 Data Requirements By Capability Tier

The Capability Tier framework defines the minimum data requirements needed for a language to support progressively more advanced forms of digital and AI-enabled interaction. These requirements are not maximum targets; they are baseline thresholds that enable specific categories of functionality.

The table below outlines the minimum linguistic, cultural, and governance requirements associated with each Capability Tier. Thresholds are indicative and should be adapted to local context, community priorities, and available resources. Progression across Tiers is not mandatory: communities may remain at a given level depending on their intended use cases, values, and governance readiness.

The requirements reflect three principles:

- Higher AI capability requires not only more data, but more structured, diverse, and contextually rich data.
- Cultural and governance requirements scale alongside technical requirements.
- Data quality, representativeness, and governance are as critical as data volume.

Category	Tier 1: Discovery	Tier 2: Interactive Tools	Tier 3: Generative AI
Primary purpose	Enable search, indexing, and basic digital presence.	Enable structured NLP tasks and interactive tools (e.g. POS tagging, NER, ASR, MT).	Enable context-aware, generative AI applications and advanced reasoning.
Text data volume	5,000–20,000 tokens.	100,000–2 million tokens.	5+ million tokens across multiple domains.
Speech data	Optional (30–60 minutes, transcribed).	3–10 hours, segmented and transcribed.	20–100+ hours, diverse, multi-speaker, context-rich.
Data diversity	Limited genres (e.g. narratives, conversations).	Multiple genres, registers, dialects, and contexts.	High diversity across genres, domains, discourse types, and sociolinguistic variation.

Category	Tier 1: Discovery	Tier 2: Interactive Tools	Tier 3: Generative AI
Annotation level	Sentence segmentation and basic metadata.	POS tagging (UD), morphology, NER, code-switching, ASR segmentation, parallel text alignment.	Semantic role labeling, discourse structure, pragmatics, sentiment, idioms, metaphors.
Cultural metadata	Basic communicative function and context tags.	Expanded cultural context (idioms, politeness, sociolinguistic variation).	Deep cultural-pragmatic annotation (ritual language, humor, indirectness, cultural norms).
Governance metadata	Basic consent, license, and provenance information clearly defined at the data record level.	Governance metadata extends to interactive and system-facing use cases: explicit consent types aligned to AI use (training, translation, ASR), clearly defined permitted and prohibited uses, provenance tracking, contributor and reviewer attribution, cultural sensitivity flags, and usage conditions for derivative outputs. Metadata must be structured and machine-readable to enable automated validation.	Fully machine-readable governance policies including consent models, licensing frameworks, permitted/prohibited AI use cases, re-consent requirements, stewardship rules, and enforcement conditions embedded into data pipelines.
Interoperability	Basic metadata structure and consistent formatting.	Use of shared standards (UD, ISO 639-3, BCP-47, JSON Schema) to ensure compatibility across tools.	Full interoperability across systems with ontology alignment and linked-data structures.
Validation & QA	Manual review and basic verification.	Automated schema validation, consistency checks, and dual technical + community review.	Continuous validation pipelines, benchmarking, governance enforcement, and lifecycle monitoring.

Category	Tier 1: Discovery	Tier 2: Interactive Tools	Tier 3: Generative AI
AI capability unlocked	Keyword search, indexing, basic discovery tools.	POS tagging, NER, basic MT, ASR, Q&A systems.	Context-aware generation, instruction tuning, conversational AI, advanced reasoning.
Risk if not achieved	Language remains digitally invisible.	Language tools are inaccurate or unreliable, limiting usability.	Language is excluded from AI systems or misrepresented in generative outputs.

Governance metadata scales alongside technical capability: from descriptive documentation at Tier 1, to machine-readable governance at Tier 2, to enforceable policy embedded in pipelines at Tier 3. The dataset records governance decisions; whether those decisions are honored downstream is a separate enforcement question.

B.2 Tier-Based Readiness Benchmarks

The following benchmarks provide indicative thresholds for assessing dataset readiness across Capability Tiers. They are not strict pass/fail criteria; they are guidance to support quality assessment and decision-making. They complement the ARI: the ARI evaluates a language's overall presence in AI ecosystems, while these benchmarks assess the internal quality and functionality of specific datasets.

Category	Tier 1: Discovery	Tier 2: Interactive Tools	Tier 3: Generative AI
Readiness level description	Foundational readiness: data sufficient for visibility, indexing, and basic digital presence.	Operational readiness: data supports functional AI tools such as tagging, translation, and transcription.	Advanced readiness: data supports context-aware, generative AI and complex reasoning tasks.
Text quality & structure	Sentence segmentation accuracy ≥ 95%; consistent formatting.	Stable annotated structure with minimal parsing errors.	Multi-domain, complex, consistently structured corpora.

Category	Tier 1: Discovery	Tier 2: Interactive Tools	Tier 3: Generative AI
Annotation quality	Basic segmentation and tagging consistency.	POS accuracy ≥ 90–95%; NER ≥ 85%.	SRL accuracy ≥ 75–85%; discourse agreement (κ ≥ 0.70).
Speech & transcription	Basic manually verified transcription.	ASR WER ≤ 30% (post-correction).	ASR WER ≤ 15–20% across diverse speakers and contexts.
Metadata completeness	≥ 95% complete metadata.	≥ 98% structured linguistic, cultural, and governance metadata.	Fully structured, machine-readable metadata across the dataset.
Cultural representation	Basic communicative-function tagging.	Idioms, politeness, and sociolinguistic variation.	Deep cultural-pragmatic representation (ritual, metaphor, context sensitivity).
Governance compliance	Consent, license, provenance present.	Full governance metadata; no conflicts in permitted/prohibited uses.	Machine-readable governance rules enforceable in AI pipelines.
Interoperability	Basic formatting consistency.	Alignment with standards (UD, ISO 639-3, BCP-47, JSON Schema).	Full interoperability including ontology integration.
Validation & QA	Manual review and spot checks.	Automated validation + dual review (technical + community).	Continuous QA pipelines with monitoring and audit logs.
Coverage & diversity	Limited genres and speaker variation.	Multiple registers, dialects, and contexts.	Broad and balanced coverage across domains and discourse types.
AI capability enabled	Search, indexing, discovery.	POS tagging, NER, MT, ASR, Q&A tools.	Instruction tuning, conversational AI, reasoning, generation.

Benchmarks should be applied contextually. For under-resourced languages, progression is incremental and adapted to community priorities. Quality, representativeness, and governance should be prioritized over strict numerical targets.

B.3 Recommended Software Toolkit

The MVD framework recommends a set of open-source tools that enable annotation, speech processing, validation, and dataset management. Selection criteria: accessibility (free or low-cost), ease of use for non-technical contributors, compatibility with open standards, and support for community ownership and control. The toolkit is a practical enabler; it does not define the rules governing the data.

Tool category	Recommended tools	Primary function	Output produced	Notes
Text & multimodal annotation	Label Studio, Doccano	Annotate linguistic data (POS, NER, classification, dialogue, instructions).	Structured annotated text (JSON, JSONL).	User-friendly; supports both technical and non-technical contributors; supports collaborative annotation.
Speech annotation & alignment	ELAN, Praat	Segment and align audio with text; analyze speech features.	Time-aligned transcripts (.eaf, TextGrid, JSON).	Widely used in linguistics; supports multi-speaker and dialect-sensitive annotation.
ASR pre-processing (draft transcription)	Whisper, Vosk	Generate initial speech-to-text transcripts for manual correction.	Draft transcripts (text + timestamps).	Outputs must always be reviewed; especially important for low-resource languages.
Parallel text alignment	LF Aligner, custom scripts	Align source-target sentence pairs for translation tasks.	Parallel sentence pairs (JSON, CSV).	Critical for Tier 2 machine translation use; alignment quality directly affects performance.
Data validation & schema checking	JSON Schema (Draft-07), custom validators	Ensure datasets conform to required structure and governance rules.	Validation reports, schema-compliant data.	Enables automated quality checks and enforcement of governance metadata.

Tool category	Recommended tools	Primary function	Output produced	Notes
Metadata & provenance management	YAML/JSON sidecar files, manifest tools	Store governance, cultural, and provenance metadata.	Structured metadata files.	Ensures traceability, accountability, and compliance with governance frameworks.
Dataset integrity & version control	Git, checksum tools (SHA-256), manifest generators	Track changes, verify data integrity, manage versions.	Versioned datasets, manifest files.	Supports reproducibility and auditability across the dataset lifecycle.
Workflow orchestration (optional, advanced)	Python scripts, CI/CD pipelines (e.g. GitHub Actions)	Automate validation, testing, and dataset release processes.	Automated validation logs and build outputs.	More relevant for Tier 3 or institutional implementations.
Collaboration & review	GitHub, GitLab, shared platforms	Enable contributor coordination, review workflows, issue tracking.	Annotated datasets, review logs.	Supports the dual review model (technical + community).

B.4 Folder Structure And Versioning Protocol

Files should be organized by Tier with subfolders for annotations, speech, ontologies, governance, and manifests. Each release should include a manifest listing file paths, byte sizes, and cryptographic checksums (e.g. SHA-256), tagged with semantic version numbers indicating whether a change is backward-compatible or breaking.

Recommended folder layout:

```
language_code/
├── tier_1/
│   ├── text/
│   ├── audio/
│   └── metadata/
├── tier_2/
│   ├── parallel_text/
│   ├── speech/
│   └── annotations/
```

```
└── tier_3/
    ├── ontologies/
    ├── discourse/
    └── governance/
```

Semantic versioning:

- v1.0.0: first release of a dataset.
- v1.1.x: small improvements (e.g. typo fixes in metadata).
- v1.2.x: new labels or annotations added.
- v2.x.x: major reorganization or new ontology layers.

Each release should expose schema.org/Dataset and/or DCAT JSON-LD so catalogs can automatically update records.

B.5 Json Examples Per Tier

Tier 1, minimal metadata record:

```
{
  "id": "tpi_t1_0001",
  "language": "tpi",
  "dialect": "tpi-PNG-Urban",
  "genre": "conversation",
  "register": "informal",
  "provenance": { "source": "community_session", "location": "Port Moresby" },
  "cultural_tags": { "function": ["storytelling"], "sociolinguistic": [], "risk": [] },
  "governance": {
    "consent_type": "community_public",
    "license": "CC BY-NC 4.0",
    "permitted_use": ["search","education"],
    "prohibited_use": ["foundation_pretraining"]
  }
}
```

Tier 2, UD POS + NER + annotation provenance:

```
{
  "doc_id": "tpi_t2_0107",
  "language": "tpi",
  "sentences": [{
    "sid": 1,
    "text": "Mi go long maket.",
    "tokens": [
```

```
      {"id": 1, "form": "Mi", "upos": "PRON"},
      {"id": 2, "form": "go", "upos": "VERB"},
      {"id": 3, "form": "long", "upos": "ADP"},
      {"id": 4, "form": "maket", "upos": "NOUN"}
    ],
    "entities": []
  }],
  "annotation_provenance": {
    "annotator_id": "ann_002",
    "reviewer_id": "rev_001",
    "guidelines": "CODI_Tier2_v1.1",
    "timestamp": "2026-02-20T11:45:00Z"
  }
}
```


Tier 3, discourse units + SRL + ontology links + governance:


```
{
  "doc_id": "tpi_t3_0204",
  "language": "tpi",
  "text": "Sapos san i go daun, yumi mas go bek long ples kwiktaim.",
  "discourse_units": [
    {"uid": "u1", "type": "CONDITION", "span": [0, 20]},
    {"uid": "u2", "type": "ADVICE", "span": [22, 61]}
  ],
  "srl": [{
    "predicate": {"form": "go", "span": [30, 32], "frame": "MOTION"},
    "arguments": [
      {"role": "ARG0", "text": "yumi", "span": [22, 26]},
      {"role": "ARG1", "text": "bek long ples", "span": [33, 46]},
      {"role": "ARGM-MNR", "text": "kwiktaim", "span": [47, 55]}
    ]
  }],
  "ontology_links": [
    {"ontology": "CODI-CPO", "class": "CommunicativeFunction:advice"}
  ],
  "governance": {
    "consent_type": "community_public",
    "license": "CC BY-NC 4.0"
  }
}
```

Appendix C. Cultural Richness Class, Full Definitions

The Cultural Richness Class is a six-point scale of how well AI output preserves cultural meaning.

C.1 The Six Classes

Class	Label	What this means in practice
Class 0	Digital Invisibility	Culture is virtually absent from the AI's outputs. Any mentions are accidental and contain factual errors about basic geography, history, or belief systems.
Class 1	Linguistic Literalism	Basic grammatical adequacy but no cultural resonance. Words are recognized literally but their weight is missed, producing comedic miscommunications or unintended linguistic microaggressions.
Class 2	"Softmaxing" Culture	Unique cultural traits are flattened into generic, Western-centric expressions. The AI relies on romanticised tropes rather than authentic, living representation.
Class 3	Cultural Contextualisation	The AI identifies topics and issues relevant to the community and can recall local common-sense knowledge such as regional holidays, food, and public figures.
Class 4	Figurative and Pragmatic Fluency	The model preserves culturally grounded meaning across complex figurative language, handling regional metaphors, idioms, and puns that reflect the community's lived experience.
Class 5	Representational Fidelity	The system achieves full Epistemic Responsibility: an accurate, context-sensitive reproduction of the group's knowledge systems, respecting high-context communication norms. Achieved through ongoing relational accountability with the community, not a fixed endpoint.

C.2 Assessment Requirements Per Class

Class 0 and Class 1 can be assessed at Level 1 (desk research). Class 2 and above require Level 2 (community self-assessment) or Level 3 (third-party audit). The reason is straightforward: the lower Classes describe failures visible from public AI outputs, while higher Classes require lived knowledge of the community to verify.

C.3 Identifying Expected-Vs-Actual Gaps

Record both the current Cultural Richness Class and the Class that would be expected given the language's ARI Class, then specify what is missing and who conducted the assessment. A language might have strong technical capacity at ARI Class 4 but still produce Class 2 (Softmaxing) outputs: a clear gap requiring intervention.

Appendix D. Cultural and Ethical Gap Analysis, Worksheet, Evidence Sources, and Assessor Questions

D.1 Per-Principle Scoring Criteria

For each CARE principle, identify Status as ✓ Present, ⚠ Weak, or ✗ Gap and cite evidence. The criteria and example evidence below give assessors a consistent rubric.

Collective Benefit (C). Does the community receive equitable benefits from AI development?

Status	Characteristics	Examples of evidence
✓ Present (Score 4–5)	Community owns resources OR receives substantial financial benefits.	Community-controlled fund documented with bank statements; revenue-sharing contract; annual report shows distributed benefits; community infrastructure built with AI project funds.
⚠ Weak (Score 2–3)	Some benefits but limited or to individuals only.	Payment to 2–3 community members as data collectors; one-time workshop provided (no ongoing benefits); free access to resulting AI tools; acknowledgment in papers but no financial return.
✗ Gap (Score 0–1)	No benefits or token acknowledgment only.	Data scraped from web without community contact; community mentioned in footnote only; commercial use but no revenue sharing; community has no idea the AI project exists.

Authority to Control (A). Does the community have decision-making power over their linguistic data?

Status	Characteristics	Examples of evidence
✓ Present (Score 4–5)	Community has veto power and can withdraw data.	Contract states community owns data; written consent protocol (community-created); community reps have voting seats on the governing committee; technical mechanism for data withdrawal exists and works.

Status	Characteristics	Examples of evidence
⚠ Weak (Score 2–3)	Consultation but no binding authority.	Community consulted but input is advisory only; representatives present at meetings but no vote; consent obtained initially but cannot be withdrawn; MOU mentions partnership but gives control to a university.
✗ Gap (Score 0–1)	No community input on decisions.	No consent protocol exists; external institution owns data license; community learned about the project from a news article; no mechanism for the community to stop unwanted uses.

Responsibility (R). Are developers accountable to the community?

Status	Characteristics	Examples of evidence
✓ Present (Score 4–5)	Robust accountability with enforcement.	Quarterly reports to community (public on website); grievance process documented with response timelines; community conducted audit (report available); documented case of community complaint being addressed; financial transparency (budget shared annually).
⚠ Weak (Score 2–3)	Some reporting but limited enforcement.	Annual email update to community (not detailed); grievance email exists but no documented responses; transparency about data sources but not finances; community can request info but no guaranteed access.
✗ Gap (Score 0–1)	No accountability to community.	No reporting to community; no grievance mechanism; community cannot access project information; developers accountable only to funders, not community.

Ethics (E). Are cultural protocols and sacred or indigenous knowledge respected?

Status	Characteristics	Examples of evidence
✓ Present (Score 4–5)	Community defines and controls cultural protocols.	Cultural protocol document (community-authored or co-authored); elder council has formal role with equal decision authority; sacred knowledge exclusion list

Status	Characteristics	Examples of evidence
		(maintained by community); use restrictions documented and technically enforced; evidence of excluding culturally inappropriate content.
⚠ Weak (Score 2–3)	Some cultural awareness but not systematic.	Generic ethics statement (not community-specific); elders consulted once but no ongoing involvement; dataset card mentions cultural sensitivity; some content excluded but no clear protocol.
✗ Gap (Score 0–1)	No consideration of cultural protocols.	No cultural protocol exists; elders and knowledge keepers not involved; sacred stories included in training data; cultural appropriation concerns raised by community.

D.2 Care Gap Analysis Worksheet Template

Language	[Language Name]	ARI Class	[0–5]

CARE Principle	Status	Gap Description
Collective Benefit	✓ / ⚠ / ✗	What is missing? What needs to be addressed?
Authority to Control	✓ / ⚠ / ✗	What is missing? What needs to be addressed?
Responsibility	✓ / ⚠ / ✗	What is missing? What needs to be addressed?
Ethics	✓ / ⚠ / ✗	What is missing? What needs to be addressed?
Cultural Representation Quality (Cultural Richness Class 0–5)	Current Class: [0–5] / Expected Class: [0–5]	Is there a gap between expected and actual Cultural Richness Class given the ARI level? Which stakeholders assessed this and how? Requires community-led assessment (minimum Level 2).

D.3 Swahili Worked Example (ARI Class 3, Some Gaps)

CARE Principle	Status	Gap Description
Collective Benefit	✓ Present	Community-led organizations receive funding. Training programs for community members. Educational resources developed.
Authority to Control	✓ Present	Community representatives on the development team. Consent protocols documented. The community has input on data-use decisions.
Responsibility	⚠ Weak	Annual reporting exists but no formal contestability process. Needs strengthened accountability structures.
Ethics	✓ Present	Cultural protocols documented and followed. Community elders consulted. Respectful approach to language preservation.
Cultural Representation Quality	⚠ Current: Class 3 / Expected: Class 3	Cultural Contextualisation is broadly achieved: the AI handles regional knowledge and common-sense cultural references accurately. However, figurative fluency (Class 4) has not yet been assessed by community members with specialist knowledge of Swahili idiom and high-context communication. This requires community-led assessment before claims of Class 4 readiness. Assessed at Level 1 (desk research) only; community self-assessment pending.

Summary: Swahili demonstrates strong technical capacity (ARI Class 3) and broadly sound governance. There is a minor accountability gap in the Responsibility dimension. On Cultural Representation Quality, the language currently meets the expected Class 3 standard, though progression to Class 4 (Figurative and Pragmatic Fluency) requires a community-led assessment before it can be confirmed. Conditionally ready for ethical deployment; community self-assessment of cultural richness should follow.

D.4 Assessor types

A Cultural and Ethical Gap Analysis may be conducted by different stakeholders with different knowledge and access. Each brings distinct strengths and limitations:

Assessor type	Strengths	Limitations
Community members	Know informal governance structures. Understand cultural protocols. Lived experience of benefit distribution.	May have conflicts of interest. May not have access to external documents (contracts, etc.).
External researchers	Can access public documents. Should have a neutral perspective. Can compare across languages.	May miss informal governance. May not understand cultural context. Cannot access private or confidential agreements.
AI developers	Know technical implementation. Have access to internal documents. Understand actual data flows.	Conflict-of-interest risk: developers may overstate governance. Community may not trust the assessment.
Third-party auditors	Independent and neutral. Can request access to confidential documents. Expertise in governance assessment.	Resource-intensive and expensive. May lack cultural context. Requires community cooperation to access information.

D.5 Evidence Sources By Accessibility

Different stakeholders have access to different evidence. The four tiers of accessibility are:

1. Public documents (accessible to external researchers): dataset documentation on Hugging Face; model cards; academic publications about the AI project; global reports (UN, OECD); community organization websites and public statements; open-source code repositories; Translation Commons "Zero to Digital" guidelines.[9]
2. Public-facing documents (require permission): annual reports from independent research organizations, NGOs, community organizations; partnership agreements via annual reports; funding reports and obtained grants or proposals; meeting minutes and agendas from council and committee meetings; newsletters, op-eds, NGO blogs.

[9] Translation Commons (2026). Zero to Digital Resources. https://translationcommons.org/zero-to-digital-resources. A step-by-step open-access resource for indigenous language digitisation covering terminology, data gathering, and font design. Developed in partnership with UNESCO's International Decade of Indigenous Languages. Particularly relevant for languages at ARI Class 0–2, where digitisation infrastructure is absent. Assessors can use these checklists to determine what digitisation steps remain outstanding for a given language.

3. Confidential documents (require agreement): contracts and revenue-sharing arrangements; data licenses; financial records of benefit distribution; internal governance structure.

4. Primary data collection (requires community participation): community interviews with key members and partners; focus groups (co-design methods); community surveys; observed governance processes in practice; data-flow audits.

D.6 Requirements By Assessment Level

- **Level 1:** Desk Research (External Researchers). Evidence: public (high-quality) documents only. Assessor: any researcher can complete this. Output: preliminary gap identification.
- **Level 2:** Community Self-Assessment. Evidence: public-facing documents plus confidential records. Assessor: community members with governance knowledge. Output: detailed gap analysis with supporting evidence.
- **Level 3:** Third-Party Audit. Evidence: all documents plus primary data collection. Assessor: independent auditor with community cooperation. Output: verified gap analysis with recommendations.

D.7 Per-Assessor Question Lists

Different assessors can answer different questions based on their relationship to the language and access to information.

Questions external researchers can answer (using public documents):

- Is the community mentioned in dataset documentation?
- Are community members listed as dataset or model creators?
- What license is used? (Community-friendly vs. restrictive?)
- Do published papers acknowledge community contribution?
- Is there public evidence of partnership agreements?

Questions community members can answer (lived experience):

- Do community members benefit from AI projects?
- How are decisions about language data actually made?
- Are elders and knowledge keepers meaningfully involved?

- Can the community stop uses they disagree with?
- Are cultural contexts and sensitivities followed in practice?
- What informal governance structures exist?

Questions AI developers can answer (technical and contractual):

- What consent mechanisms are technically implemented?
- Can the community actually withdraw data? How?
- What financial agreements are in place?
- How is sacred knowledge identified and excluded?
- What accountability mechanisms exist?

Questions third-party auditors can answer (comprehensive verification):

- Do formal agreements match practice?
- Are benefits actually reaching the community?
- Do consent processes function as documented?
- Are contestability mechanisms effective or symbolic?
- Where are the gaps between policy and implementation?

D.8 Multi-Stakeholder Workflow

For the most complete assessment, combine perspectives:

1. Step 1. External researcher conducts desk research using public documents (identifies obvious gaps).
2. Step 2. Community members complete a self-assessment questionnaire (provide insider perspective and access to semi-public documents).
3. Step 3. External verification checks that claimed governance structures actually exist (verify without judging quality).
4. Step 4 (optional). Third-party audit for high-stakes projects or when disputes exist (comprehensive verification).

Community members should lead the assessment; external researchers provide support and verification, not the reverse.

D.9 Using The Cultural And Ethical Gap Analysis Results

The Cultural and Ethical Gap Analysis surfaces three operational outputs:

1. **Readiness for deployment.** Languages with mostly ✓ Present indicators are ready for ethical deployment. Languages with ⚠ Weak indicators need to strengthen specific areas before deployment. Languages with ✗ Gap indicators are not ready: deployment would be extractive; governance must be established first.
2. **Priority areas for support.** Identifies exactly what needs to be addressed (e.g. create consent protocols, establish revenue-sharing, develop grievance mechanisms).
3. **Resource allocation.** Distinguishes between languages needing (a) technical resources, (b) governance support, or (c) both.

D.10 Relationship To The ARI

The Cultural and Ethical Gap Analysis complements the ARI. The ARI measures technical capacity; the gap analysis identifies cultural and governance gaps. Together they answer:

- What AI infrastructure exists? (ARI)
- What governance is missing? (Cultural and Ethical Gap Analysis)
- Is this language ready for ethical AI deployment? (Both)

The Cultural and Ethical Gap Analysis is performed at each ARI Class. It is not a separate score; it is a diagnostic applied to whatever technical position the language currently occupies.

Appendix E. Technical Implementation

This section gives the technical specification a community or supporting institution needs to actually build an MVD-compliant dataset. The forkable starting point is the CODI-MVD GitHub repository, which provides the Tier 1–3 JSON schemas, the CODI Cultural-Pragmatic and Governance Ontologies, validation tooling, and reference samples.

The framework prioritizes four properties for the labels and metadata attached to language data:

- **Cultural Fidelity.** Labels accurately reflect the culture.
- **Provenance.** A transparent record of where the data came from and who reviewed it.
- **Community Control.** The community decides what categories and tags are used.
- **System Compatibility.** Labels work across different digital platforms.

E.1 Tier Data Requirements

Each Capability Tier (§3.2) requires a defined volume of text and speech, structured annotations of increasing depth, cultural and governance metadata, and quality gates that must be cleared before promotion. In summary:

- **Tier 1 (Basic Discoverability)**: a few thousand tokens, sentence-segmented, with linguistic + cultural + governance metadata published in machine-readable form.
- **Tier 2 (Interactive Tools)**: larger text and speech volumes with POS tags, morphological features, named entities, code-switching flags, parallel sentences, and ASR-ready speech segments.
- **Tier 3 (Generative AI)**: semantic role labeling, discourse structure, idioms and metaphors with glosses, instruction-tuning pairs, and links to lexical-semantic ontologies, accompanied by machine-readable governance policy files.

The full per-tier specification (token thresholds, audio hours, coverage requirements, inter-annotator agreement (IAA) targets, NER span-F1 thresholds, ASR word-error-rate bounds, governance metadata requirements, structural requirements, and quality gates) is in Appendix B.

E.2 Metadata Schema: Three Layers

Every record carries three layers of metadata:

- **Layer 1: Linguistic Basics**. Language, dialect, and script identifiers (ISO 639-3, BCP-47); genre, register, speaker roles, and source type.

- **Layer 2: Cultural Context**. Communicative function (e.g., storytelling, blessing, apology); sociolinguistic context (e.g., youth or diaspora registers); politeness and honorifics; explicit risk flags for sacred, ceremonial, or otherwise sensitive material.
- **Layer 3: Governance**. Consent type, license, permitted and prohibited uses, stewardship roles, and at higher tiers re-consent intervals.

Dataset-level documentation should be published using the Datasheets for Datasets format (Gebru et al. 2018) and, where helpful, a Dataset Nutrition Label to surface risks, limitations, and intended uses before data is applied.

Validation rules that apply at every tier:

- Consent and license fields are mandatory.
- If cultural_tags.risk contains sacred or ceremonial, access must be restricted and governed by the community steward.
- Code-switching records must specify the target language code.

A minimal governance-heavy record looks like this:

```
{
  "id": "lg_t2_0421",
  "language": "lug",
  "genre": "narrative",
  "register": "formal",
  "cultural_tags": { "function": ["storytelling"], "risk": ["community_sensitive"] },
  "governance": {
    "consent_type": "community_public",
    "license": "CC BY-NC 4.0",
    "permitted_use": ["education", "research_non_commercial"],
    "prohibited_use": ["open_commercial_release_without_royalty"]
  }
}
```

Full per-tier JSON examples are in Appendix B.

E.3 Annotation Depth By Tier

Annotation depth grows with tier:

- **Level 1 (basic).** Sentence segmentation, source type (oral transcription vs. written), and core cultural and governance metadata.

- **Level 2 (intermediate).** Universal Dependencies POS tags and morphological features, named entities, code-switching flags with target language tags, parallel sentence alignment, and speech segments with timestamps suitable for ASR training.
- **Level 3 (advanced).** Semantic role labeling, discourse units (e.g., condition, advice), idioms and metaphors with literal glosses, and links to OntoLex-Lemon entries so lexicon and ontology stay synchronized.

E.4 Label Sets And Ontologies

The framework defines three CODI-managed dictionaries:

- **Language Dictionary.** Grammar and morphological structure (Universal Dependencies UPOS + features).
- **Cultural-Pragmatic Ontology (CODI-CPO).** High-level categories such as Sacred Knowledge, Social Roles (e.g., Elder, Ritual Specialist), and Knowledge Types (e.g., medicinal, seasonal).
- **Governance Ontology (CODI-GOV).** Authority over the data and specific usage rights.

For cross-lingual interoperability the MVD adopts Universal Dependencies' 17 Universal POS tag categories and its morphological feature inventory, which leverages hundreds of existing treebanks and tools while allowing local XPOS extensions where necessary. For lexical and semantic relations, OntoLex-Lemon provides a W3C community-report model supporting dictionary-as-linked-data applications, including variation and translation modules useful in multilingual and low-resource contexts.

Both CODI-CPO and CODI-GOV are published as JSON in the GitHub repository and are intended to be extended by communities over time.

E.5 Quality Assurance And Intake

Quality assurance is a continuous practice rather than a final check. The intake pipeline has four gates:

1. **Automated schema check**. Files land in an incoming area with temporary IDs. A validator scans for missing required fields, inconsistent dialect tags, and structural errors.
2. **Governance check**. Verifies consent is present and that no file lists the same use as both permitted and prohibited. Rejected files are returned with a reason.
3. **Checksum check**. Verifies cryptographic signatures (SHA-256) to confirm files were not corrupted in transit.

4. **Dual review**. Every file passes a Technical Reviewer (digital structure) and a Community Reviewer (cultural accuracy and appropriateness) before promotion to the stable library.

Annotators and reviewers are trained through a three-stage program covering ethics and governance (CARE and OCAP; see §5), language basics (POS, NER, segmentation), and advanced skills (semantic role labeling, discourse, ontology linking) for Tier 3 work. Calibration exercises ensure that two annotators given the same input produce the same labels.

E.5.1 Privacy and cultural safety

Data is categorized into three access levels:

- **Public**. Safe for everyone.
- **Restricted**. Requires specific permission from a community steward.
- **Ceremonial-Only**. Accessible only to designated cultural authorities.

Data minimization applies throughout: only the information necessary for the project is collected. Material identified as sacred or restricted is quarantined by default. For Tier 3 generative use, the framework recommends a 24-month re-consent interval, encoded in machine-readable policy files consistent with CARE's emphasis on authority and ethics and OCAP®'s assertion of community control.

E.5.2 Benchmarks per tier

Per-tier readiness benchmarks (text quality, annotation IAA, ASR WER, metadata completeness, cultural representation, governance compliance, interoperability) are in Appendix B. These are guidance for assessing dataset readiness, not strict pass/fail criteria. Progression depends on community priorities and on the maturity of governance and cultural representation, not on hitting numerical targets in isolation.

To communicate readiness status across partners and communities, datasets can be summarized using a Green / Amber / Red scorecard against the Tier benchmarks above. This gives a quick visual snapshot of dataset health and where additional resources are needed.

E.6 Open Standards

The framework uses open, widely adopted standards so datasets remain portable, future-proof, and interoperable with global tools.

Focus area	Standard	Purpose
Language codes	ISO 639-3 (base codes)	Unambiguous language, dialect, script, and region identification.

Focus area	Standard	Purpose
	BCP-47 / RFC 5646 (subtags)	
Text encoding	UTF-8	Universal character encoding.
Data exchange	JSON / JSONL	Primary record format.
Long-form textual encoding	TEI P5	Structural, renditional, and semantic features for edition-level material.
Grammatical annotation	Universal Dependencies + CoNLL-U	Consistent POS and feature annotation across languages, with local XPOS extensions allowed.
Lexical / semantic linking	OntoLex-Lemon (W3C)	Dictionary-as-linked-data.
Schema validation	JSON Schema Draft-07	Deterministic validation of structure and governance fields.
Speech audio	WAV PCM mono 16 kHz	Standard ASR input format.
Speech alignment	ELAN (.eaf), Praat (.TextGrid)	Time-aligned transcription, convertible to MVD JSON segment structures.
Licensing	SPDX identifiers	Standardized license labels (e.g., CC-BY-NC-4.0).
Catalog metadata	schema.org/Dataset + W3C DCAT	Federated discovery via web search and data portals.

Using these standards rather than proprietary alternatives guards against vendor lock-in, keeps community-defined governance signals permanently attached to the data, and allows communities to iterate locally while remaining compatible with the wider digital world.

E.7 The Reference Implementation

The CODI-MVD GitHub repository provides a forkable starting point:

- JSON Schema Draft-07 files for Tier 1, 2, and 3 metadata.

- The CODI-CPO (Cultural-Pragmatic) and CODI-GOV (Governance) ontologies as JSON.
- validate_item.py and validate_repo.py for per-record and bulk schema validation.
- checksum_manifest.py for manifest generation and SHA-256 verification.
- A CI workflow for automated pull-request validation.
- Reference sample folders per tier (currently scaffolded; populating with real data is part of Phase 2).
- Documentation stubs for the contributor toolkit, dataset cards, governance policy, and a crosswalk to external standards.

Communities and supporting institutions implementing the MVD are encouraged to fork this repository rather than start from scratch.

A note on enforcement: the metadata fields the MVD specifies (consent type, license, permitted and prohibited uses, stewardship roles) bind downstream behavior only insofar as the tools and people handling the data honor them. The MVD documents the conditions of ethical use. It is up to those that build, use, and maintain the data and AI system to ensure that reality follows the conditions. Real-world enforcement of data terms across jurisdictions and platforms remains an open problem and is outside this framework's scope though it is a likely subject of Phase II work.

Appendix F. Working Group Methodology

F.1 Project Inception

The MVD project was initiated under the broader mission of the CODI organization, which seeks to ensure that speakers of all languages, particularly Indigenous, minority, and oral-tradition languages, can participate meaningfully in the digital and AI ecosystem. Recognizing that many languages remain underrepresented in digital systems due to the lack of structured and culturally appropriate datasets, CODI launched the Minimum Viable Dataset project as its first foundational effort.

The objective is to define the minimum dataset requirements necessary for languages and communities to participate in today's digital and AI-driven environment. The work focuses on identifying dataset thresholds, establishing readiness benchmarks for different levels of digital use, and developing practical pathways that enable communities to build and strengthen their linguistic data resources.

F.2 Intended Audience

This report is intended for a range of stakeholders involved in language preservation, digital inclusion, and the development of AI and digital technologies. The primary audience includes:

- Language communities, particularly those representing Indigenous, minority, and oral-tradition languages, seeking practical guidance on structuring, developing, and managing datasets that enable their languages to participate meaningfully in the digital and AI ecosystem.
- Policymakers and public-sector institutions responsible for shaping national and international digital strategies, cultural-preservation initiatives, and language-inclusion policies.
- Researchers, linguists, and academic institutions working in computational linguistics, language documentation, AI, and digital humanities.
- Technology developers and AI practitioners, including companies, startups, and open-source communities building digital tools, language technologies, and AI systems.
- Funders, philanthropic organizations, and development partners seeking to support initiatives that expand digital access and representation for underserved languages.
- Civil society organizations, cultural heritage institutions, and community-based organizations engaged in language preservation, cultural documentation, and digital empowerment.

F.3 Working Group Formation

To guide the development of this framework, CODI established a multi-stakeholder Working Group composed of experts and practitioners from linguistics, AI, cultural heritage, data governance, and community-based knowledge systems. The Working Group was tasked with developing recommendations, technical considerations, and governance principles to inform the definition and implementation of the MVD framework. See https://www.codi.global/newsroom/codi-assembles-global-experts-to-power-the-next-era-of-multilingual-ai for the full Working Group composition.

F.4 Working Methods

The Working Group adopted a collaborative and structured working process combining regular meetings, expert presentations, written contributions, and iterative drafting. As of the preparation of this report, the Working Group has convened formal meetings to review project objectives, examine relevant frameworks and practices, and develop preliminary recommendations and implementation guidance. Discussions were conducted through open dialogue among members, supported by background research and shared documentation. The Working Group emphasized a multidisciplinary perspective to ensure that the resulting framework reflects both technical requirements and community-centered cultural considerations.

F.5 Readings Process

To support iterative development and review of the proposed framework, the Working Group conducted three formal readings of the draft materials.

- First Reading. Focused on introducing and discussing the foundational concepts of the MVD framework, including the scope of the project, core definitions, and initial structural proposals. Members reviewed early draft components and provided feedback to refine the direction of the work.
- Second Reading. Examined the updated draft in detail, incorporating revisions from earlier discussions. Members reviewed the evolving recommendations, assessed alignment with the project's objectives, and proposed additional refinements.
- Third Reading. Produced the present final report. Members reviewed the consolidated framework, addressed remaining editorial questions, and finalized the recommendations.

F.6 Work Streams (Tracks)

To organize the work effectively and address the interdisciplinary nature of the project, the Working Group structured its efforts around two primary tracks.

- **Track 1:** Cultural, Ethical, and Sustainability. Examines the governance and sustainability aspects of the MVD framework. Work includes economic and resource considerations for dataset development, sustainable models that enable communities to create and maintain datasets, and relevant ethical standards. Particular attention is given to cultural respect, community ownership, informed consent, data sovereignty, and responsible data stewardship.

- **Track 2:** Language Datasets, Technical Requirements, and Other Technical Considerations. Focuses on the technical aspects required for dataset development and usability within digital and AI systems. Areas of work include approaches to data labeling and annotation, linguistic and cultural metadata structures, benchmarks per Tier, technical interoperability, and other practical considerations necessary for integrating datasets into digital platforms and AI applications.

F.7 Methodological Principles

Across both tracks, the Working Group's methodology is guided by four core principles:

- **Inclusivity.** Ensuring that the framework reflects the needs and realities of diverse linguistic communities.

- **Interdisciplinary collaboration.** Integrating expertise from technical, linguistic, cultural, and governance perspectives.

- **Iterative development.** Refining proposals through structured readings, discussions, and continuous feedback.

- **Community-centered design.** Recognizing that linguistic and cultural data must be developed in partnership with and guided by the communities to whom the languages belong.

https://www.codi.global/
info@codi.global