INTRODUCING THE DATA TO AI VALUE CHAIN

PROMOTING LINGUISTIC AND CULTURAL DIVERSITY FOR AI DIFFUSION



A JOINT PUBLICATION WITH CURRENT AI

September 2026

Copyright © UNDP 2026. All rights reserved.

One United Nations Plaza, NEW YORK, NY10017, USA.

UNDP is the leading United Nations organization fighting to end the injustice of poverty, inequality, and climate change. Working with our broad network of experts and partners in 170 countries, we help nations to build integrated, lasting solutions for people and planet. Learn more at undp.org or follow @UNDP @UNDPDigital.

The views expressed in this brief are those of the author(s) and do not necessarily represent those of the United Nations, including UNDP, or the UN Member States. UNDP does not endorse any entity, brand, product, or service.

PREFACE

Artificial Intelligence (AI) research is advancing and so are the multilingual capabilities of AI models. The volume of linguistically diverse data being collected is increasing, and new tools for translation, voice recognition and other functions are being released at a fast pace. So why do so few Language AI initiatives reach the communities that they are meant to serve? The gap between what AI can do in evaluations and sandboxes and what it does for communities speaking underrepresented languages is not a technical accident. It is the product of the current alignment of investment, infrastructure, and incentives, which continue to underestimate the communities that stand to benefit the most from public interest AI.

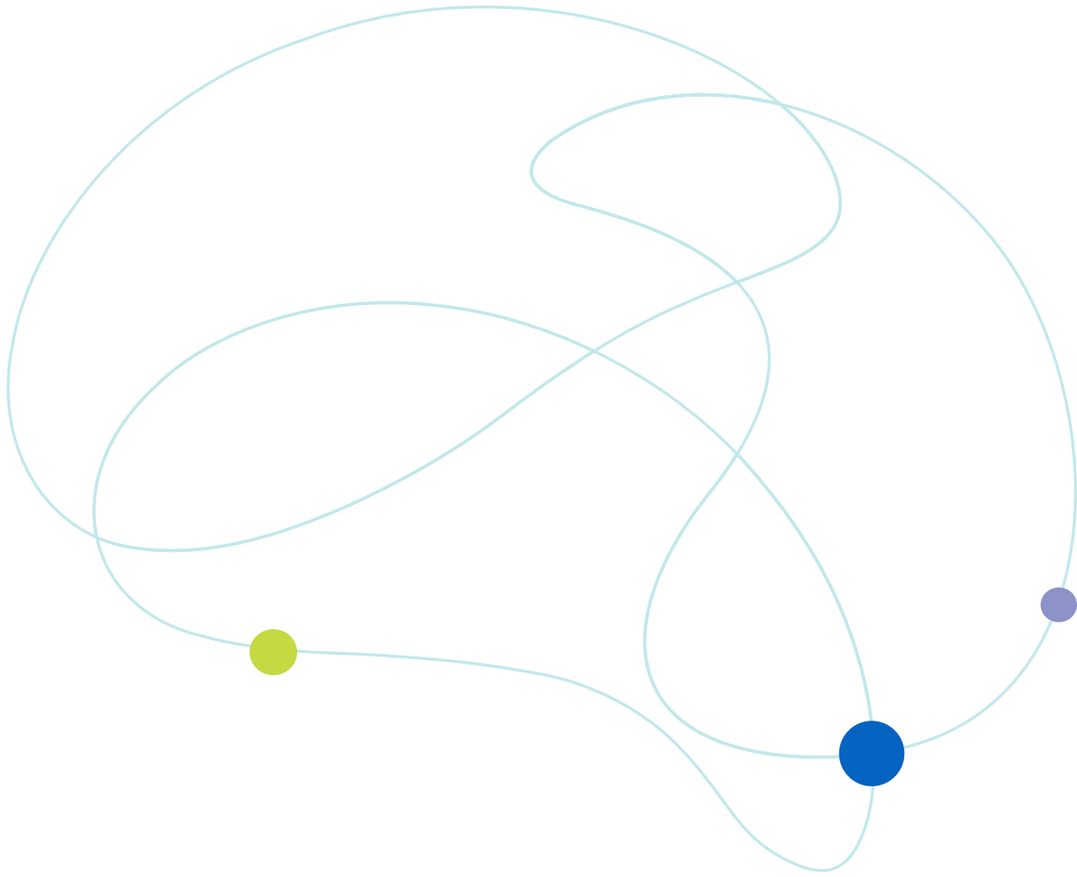
AI is diffusing into economies and institutions, but diffusion alone is not a development outcome. It is a signal that certain enabling conditions are already in place. This brief interrogates whether this diffusion is happening on terms that build agency: whether countries retain the institutional capability to understand, negotiate, supervise and correct the systems diffusing into their public and economic life, and whether that diffusion builds and retains capability and value locally rather than passing through. The Data-to-AI Value Chain for linguistic and cultural diversity is a core feature of AI agency.

This brief is offered in a spirit of public interest. It asks: what it would take for AI to work for the languages and markets that already attract capital and research attention? And, what it would take for AI to serve the billions of people who speak underrepresented local languages, embody local and Indigenous cultures, and safeguard our traditional systems of knowledge? Public interest AI is a guiding principle and a design imperative: it means treating language data as a public asset to be stewarded rather than extracted, building sustainable infrastructure that more than one project can use, and measuring success by whether communities benefit, not by abstract tests of capability.

Linguistic and cultural diversity are critical to meaningful adoption that unlocks AI's collective benefit. A model can diffuse widely and be downloaded, cited, or deployed in a demo, without ever reaching the small business owner, the health worker, or the student it was meant to serve. Genuine adoption requires the slower, less visible work of governance, procurement, deployment pathways, and institutional ownership that this brief sets out to describe. In other words, scaling Language AI is not simply a matter of scaling data or compute capacity. It is a matter of scaling the conditions under which adoption becomes possible and doing so on terms that are fair to the communities whose knowledge, voices, and languages make the underlying data possible in the first place.

Fair exchange sits at the centre of that proposition. Communities that contribute their voices, stories, and cultural knowledge to Language AI initiatives are rarely credited, and even more rarely compensated, when that data goes on to create commercial value. Any theory of scale that does not also account for how value returns to its source is at best incomplete and at worst a continuation of long-standing extractive patterns under a new name. This brief argues that closing the loop between contribution and benefit is not a peripheral concern but a precondition for AI that is genuinely public interest-oriented, open, inclusive, and sustainable at scale.

To illustrate how value is created in AI innovation, this brief introduces the **Data-to-AI Value Chain**, UNDP's analytical framework for understanding where inclusive Language AI stalls between data and deployment. The framework draws on lessons from the wider public interest AI movement. This brief is a product of a collaboration between UNDP and Current AI, a non-profit international organization that funds, builds, and connects public interest AI innovation for cultural preservation and AI sovereignty around the globe. This brief consolidates these insights together with two years of implementation experience through the UNDP Local Language Accelerator and related country engagements. Building on these insights, it introduces **Every Language Matters** (ELM), a global-to-local initiative that turns those lessons into coordinated public action.



ACKNOWLEDGEMENTS

This brief was prepared by the United Nations Development Programme (UNDP) and Current AI, under the strategic direction of Robert Opp, Chief Digital Officer & Director of the Digital, AI and Innovation Hub, and Keyzom Ngodup Massally, Director, AI Hub for Sustainable Development and Head of Digital and AI Programmes at UNDP.

The brief was authored by Keyzom Ngodup Massally, Alena Klatte, Navya Alam, Mark Irura, Barbora Bromová and Ottavia Brussino.

Insights in this brief are generated from the implementation work in countries and with partners, made possible by the generous support of the Government of France, the AI Hub for Sustainable Development, UNDP AI Sprint, and the Government of Japan through UNDP Digital X. Their catalytic financial contributions and in-kind resources have been instrumental in enabling the design of UNDP's Local Language Accelerator work in countries and communities for public interest interventions on linguistic and cultural diversity.

We extend our sincere thanks to the government ministries and agencies, communities, researchers, technical partners, and practitioners in Democratic Republic of the Congo, Ghana, Iraq, Lesotho, Malawi, Mexico, Namibia, Nepal, Peru, Republic of Congo, and Serbia who have collaborated with the team over the past two years. Their insights and experiences have been instrumental in shaping the reflections presented in this brief.

We also thank Dwayne Carruthers, Alexander Hradecky, Marc Crouzat and Romilly Golding for their editorial direction and guidance throughout the development of this brief. We are also grateful to Dr. David Song, Draško Drašković, and Martin Benjamin for their thoughtful peer review and valuable feedback.

ABOUT UNDP LOCAL LANGUAGE ACCELERATORS

Since 2024, UNDP's Local Language Accelerator (LLA) has united governments, universities, technical partners, startups, and communities across eleven countries in Africa, Asia, Europe, Latin America, and the Arab region to build inclusive Language AI. The programme supports the creation of high-quality language datasets, strengthens local AI ecosystems, pilots community-driven approaches to data governance, develops public-interest use cases, and builds partnerships that move language technologies from research into practical use.

As the programme's network expanded, a shared set of challenges came into focus that went well beyond collecting and curating language data. Communities and policymakers also needed support to build responsible data governance frameworks, train and evaluate linguistically and culturally inclusive models, deploy them to solve real problems, and establish governance arrangements that ensure communities benefit from the AI systems they help create. Progress at one stage often depended on decisions made much earlier, while weaknesses at another stage limited outcomes across the rest of the system. UNDP came to understand this progression as the Data-to-AI Value Chain: the sequence of technical, institutional, financial, and governance activities through which local languages, cultures, ideas and knowledge systems take the form of data, and are transformed into digital artefacts, including AI, that create public value.

ABOUT CURRENT AI

Current AI is a global partnership of governments, philanthropists and industry leaders, founded at the AI Action Summit in Paris in February 2025. It exists to build open, public technology infrastructure that puts people over profit, coordinating and amplifying public-interest AI efforts that are otherwise fragmented across the world.

TABLE OF CONTENTS

Preface 4
Acknowledgements 6
Executive Summary 8
Introduction: Language as AI's Global Last-Mile Problem 14
Why Responsible Diffusion Remains a Challenge 16
A Global Shift in Mindset: Toward Sustainable AI Ecosystems 20
Leveraging the Data to AI Value Chain for Systematic Change: A Government's View 26
Introducing Every Language Matters: a Global Initiative to Build a Shared Vision of Inclusive Language AI 29
Linguistic and Cultural Diversity as the New AI Frontier 32
Conclusion: Let's Build Public, Inclusive AI Together 33
References 35

EXECUTIVE SUMMARY

This brief argues that inclusive, multilingual AI is not simply a technical challenge but a value-chain and diffusion problem. Progress at any stage, from data collection to deployment, depends on decisions made at every other stage, and whether the resulting knowledge travels to the next country rather than being rebuilt from scratch each time.

AI is diffusing into economies and institutions, but diffusion alone is not a development outcome. It is a signal that certain enabling conditions are already in place. This brief interrogates whether this diffusion is happening on terms that build agency: whether countries retain the institutional capability to understand, negotiate, supervise and correct the systems diffusing into their public and economic life, and whether that diffusion builds and retains capability and value locally rather than passing through.

The Data-to-AI Value Chain is a core determinant of agency for linguistic and cultural diversity. Excluding a language at any stage limits what the resulting systems can understand, serve, or be corrected by. Drawing on two years of implementation experience across UNDP's Local Language Accelerators (LLAs), this brief outlines the Data-to-AI Value Chain and identifies consequential break points. To address these, the brief introduces Every Language Matters (ELM), the global-to-local initiative that puts this value chain to work for inclusive, sustainable, public interest AI.

WHY THIS MATTERS FOR DEVELOPMENT

The question of linguistic inclusion in AI is ultimately a question of who gets to participate in and benefit from the AI-driven transformation of economies and societies. AI has the potential to accelerate sustainable development, expand access to services and opportunities, and advance progress towards the Sustainable Development Goals. But its benefits will not be guaranteed. If the systems being built work better for some languages, communities and markets than others, AI risks reproducing - and potentially widening - existing inequalities.

This is particularly consequential for communities whose languages remain under-represented in digital systems. When AI does not work effectively in the languages people use to access healthcare, education, finance, public services and markets, linguistic exclusion can become a barrier to meaningful participation and impose a ceiling on the benefits gained from new innovations. Therefore, fostering linguistic diversity is a part of the broader challenge of ensuring that AI contributes to inclusive development and leaves no one behind.

For UNDP, linguistic and cultural diversity in AI is not an end in and of itself. Rather, it is a condition for making AI work for development: for creating jobs, unlocking investment, building resilience, strengthening institutions, and enabling transparent, participative governance.

AI is diffusing into economies and institutions, which signals that certain enabling conditions are already in place. Yet it is still uncertain whether diffusion is happening on terms that build agency: whether countries retain the capability to understand, negotiate, supervise and correct the systems entering their public and economic life, and whether public value can be built and retained locally rather than passed through or extracted out.

Ensuring AI diffusion is an effective accelerator for sustainable development and public value creation requires considering the full Data-to-AI Value Chain. The value chain encompasses data collection and curation, models that perform in real linguistic conditions, assurance and trust, and ultimately deployment and adoption at scale. A country that produces language data but does not have the capacity to build, test, assure or deploy on top of it has contributed to the value chain without fully benefiting from it. The opportunity is to build the capabilities and partnerships needed to help realise AI's full value potential, and carry promising pilots through to real-world use - creating new markets, strengthening resilience, and expanding opportunities for communities and businesses.

The choices that countries make now will shape who benefits from AI. They will influence jobs, investments, market resilience, public services and safeguards, as well as determine who creates value and who captures it. Both the Data-to-AI Value Chain and Every Language Matters are designed to equip government and communities with the analytical frameworks, policy tools, and diffusion pathways needed to shape, deploy, and benefit from AI on their own terms. This shifts the focus from whether a language can be represented in a model to whether the conditions exist for AI to become a meaningful and sustainable development tool for the people and communities it is intended to serve.

KEY FINDINGS AND TAKEAWAYS:

1

The gap between well-resourced and under-resourced languages is widening: languages with more data attract more research, more research produces more models, and more models attract more data. This leaves the majority of the world's roughly 7,000 languages – and the communities that speak these languages – structurally underrepresented in AI research and development.

2

Language AI moves through five interdependent stages: data collection, curation, model training, deployment, and impact and reuse. A break at any stage limits what the rest of the chain can deliver. Current investment concentrates heavily on stage 1 – data collection – while curation standards, shared compute and evaluation infrastructure, deployment budgets, and benefit-sharing remain comparatively under resourced.

3

National AI strategies – regardless of a country's income level or institutional maturity – often stall in the same five places of the value chain. Stalling results in fragmented ministerial mandates, locked public data, unreformed procurement, missing deployment pathways, and the absence of an accountability mechanism.

4

Fair exchange, ensuring that value created from community-contributed language data returns to the communities and countries that produced it, is a precondition for sustainable scale, not an add-on to be addressed later.

5

Access does not necessarily equal adoption. Closing this gap requires treating each community's experience as a reusable pathway, with a documented playbook and toolkit, so that future adopters begin further ahead rather than starting from zero. This reflects the spirit of the wider public interest AI movement this brief draws on, including Current AI's Potluck and the 100 Pathways to AI Diffusion initiative.

HOW THESE FINDINGS ARE BEING PUT INTO PRACTICE:

- The Data-to-AI Value Chain is not only a diagnostic tool; it is an important, organizing part of the global ELM initiative. It is the shared reference map that anchors ELM's partnerships, sequences its financing asks, and structures its diffusion goals across five commitment areas: data and community sovereignty, AI infrastructure and model commons, ecosystem and capacity-building, public service deployment, and measurement and accountability.
- Modular entry allows governments to join the Every Language Matters initiative at whatever stage that matches their readiness, while working towards a shared destination: a Language AI ecosystem that is inclusive by design, sovereign in governance, and self-sustaining beyond any single funding cycle.

THE VALUE CHAIN WITHIN THE EVERY LANGUAGE MATTERS GLOBAL INITIATIVE

Within the global Every Language Matters initiative, the Data-to-AI Value Chain functions as connective infrastructure, not only a diagnostic tool. Partners can use it to target funding to the stages the market underserves, echoing Current AI's logic of funding what the market will not, most visibly stages 1 through 3: data collection, curation, and model training. Technology and infrastructure partners use it to prioritize shared compute, model commons, and evaluation standards at stages 2 and 3, so that no single country or entity need to build that infrastructure alone. Governments use it to define their local commitments across all five ELM commitment areas, each mapped to specific stages of the chain. Partnerships in ELM form around specific stages and hand-offs in the chain rather than the ecosystem. Financing is sequenced against the full chain instead of concentrated at data collection alone. Diffusion goals are measured by how quickly the whole chain moves from data to deployed public value, and by how much further ahead each new country and community can start because of it.

Every Language Matters is a global initiative designed to support that ecosystem. Building on implementation experience from UNDP's Local Language Accelerator, it provides a platform for governments, funders, researchers, technology partners, and communities to collaborate in strengthening the institutional, technical, financial, and governance foundations needed to build inclusive Language AI.

Data Collection

Safeguarding community data sovereignty

- FPIC protocols as the national standard for all publicly funded language data collection

Data Curation

Building data infrastructure commons

- Common metadata and interoperability standards for language datasets and models

Model Training

Strengthening ecosystem support

- Coordinated advocacy and funding for accessible and ‘local’ owned infrastructure
- Robust evaluation benchmarks

Deployment

Facilitating public sector capacity-building

- Local language AI tools for government services
- Support for local innovators to build and scale
- Governance both during and post-deployment of (agentic)AI

Impact and Reuse

Enabling trusted accountability

- Local feedback and impact measurement
- Community-centric value return and benefit-sharing

Continuous Feedback Loops

INTRODUCTION: **LANGUAGE AS AI'S GLOBAL LAST-MILE PROBLEM**

Despite considerable advancements in AI in recent years, the cultural and linguistic capabilities of frontier models still fall short. Landmark research shows a worrying trend: while well-resourced languages such as English, Mandarin and German dominate AI development[i], the majority of the world's languages have little to no online presence, few corpora and almost no research dedicated to them[ii]. This is the case for most African languages, along with many languages across South and Southeast Asia, the Pacific, and languages of Indigenous communities in the Americas and Australia.

The most important finding from this landmark research was that the digital inequity between languages tends to grow wider over time, instead of shrinking. Languages with more data attract more researchers; more researchers produce more models; and more models attract more data. Languages that are already well-served keep pulling ahead, while the rest fall further behind. English has reached a state of over-resourcing, gaining up to 50,000 new models a year, while languages such as Congolese-Swahili, Wolaytta (Ethiopia) and Kwangali as well as Kuanyama (Namibia, Angola), among others gain an average of only 4.15 new models per year[iii]. And when AI developers describe their systems as 'language-neutral' or 'language-agnostic', that framing does not change the reality: a model trained mostly on English and a handful of European languages will work well for those languages and contexts while performing poorly for everyone else[iv].

The picture within the African continent makes this concrete. During a review of 884 research papers on African Language AI published over five years, the volume of work quadrupled, reaching over 200 papers in 2023[v]. That is genuine progress. But more than 70 percent of the African languages covered appeared in fewer than five papers, and a small group of languages, such as Swahili, Hausa, Yoruba, and Amharic, received the bulk of attention[vi]. Africa is home to more than 2,000 languages, with at least 75 spoken by over a million people, each with distinct dialects, signifiers, traditions, and customs that remain largely absent from AI research[vii]. As a result, AI systems that mediate access to healthcare, legal services, education, and economic opportunities are not built for most of the world's people, the way they speak or understand the world. With over 7,000 languages and many more distinct cultures on Earth, this represents a profound missed opportunity.

An under-resourced language is one that lacks the digital data and technical infrastructure needed to build effective AI, not because of any property of the language itself, but because of historical patterns of under-investment and limited research focus[viii].

Yet the challenge that prevents multilingual, multicultural AI from harnessing its full potential is systemic: when language data gets collected, cleaned and curated by communities, it often takes a considerable amount of time to convert into usable tools. Local innovators often struggle to scale their work and get their products into the hands of the communities they were intended to serve. Many models are trained but never reach the users who need them. Pilots end when the grant ends[ix]. **As for the communities who contributed their voices, stories, and cultural knowledge towards these data collection efforts, they wait far too long to benefit from new models and applications and are often not credited for their contributions once it is time to collect revenue**.

Many of the institutions and pathways that are necessary for moving Language AI from pilot projects to public deployment remain underdeveloped. Language data, where it exists, is often dispersed across ministries, universities, civil society organizations, broadcasters, and local communities – with limited incentives or mechanisms to share it. Further, drawing on learnings from UNDP's Artificial Intelligence Landscape Assessments completed between 2024 and 2026, procurement systems rarely create demand for locally developed language technologies[x]. Successful pilots frequently lack an institutional home capable of carrying them into sustainable projects in the long run[xi]. As a result, the decisions about what gets built, how it gets deployed, and who profits from it are being made elsewhere, by larger actors with more resources and less connection to the communities the technology was designed to serve.

Consider what this means in practice. A small business owner in Kisumu, Kenya cannot access a loan application interface or customer service tool in Dholuo because one is yet to be built, and the commercial logic of dominant AI systems does not make it worth building, even if it is feasible. In addition, the underlying infrastructure works against it: because AI models process words in units called tokens, a word that costs one token in English may cost four or five in Dholuo, making every API call more expensive – and weakening the business case for inclusive AI. The cost of being under-resourced is therefore not just the absence of a tool, but the structured indirect costs of building one.

The consequences are dire and extend beyond individual projects. Valuable language resources remain underdeveloped and, at times, underutilized. Communities that contribute to their efforts often see little to no return from the systems they help create. The economic and social value generated from language resources increasingly flows towards larger actors[xii] with greater technical and financial capacity, rather than to the countries and communities that made those technologies possible[xiii].

WHY RESPONSIBLE **DIFFUSION REMAINS A CHALLENGE**

Since 2024, UNDP has been designing, implementing, and evaluating local interventions that address these questions with and for local language communities. Through a global network of Local Language Accelerators, 11 UNDP teams have been supporting the collection or curation of new language data, addressing prevailing representation gaps, and building or testing inclusive AI tools for optical character recognition, community-validated voice data collection, and the creation of hyperlocal dictionaries across text and voice.

UNDP Local Language Accelerators at a Glance

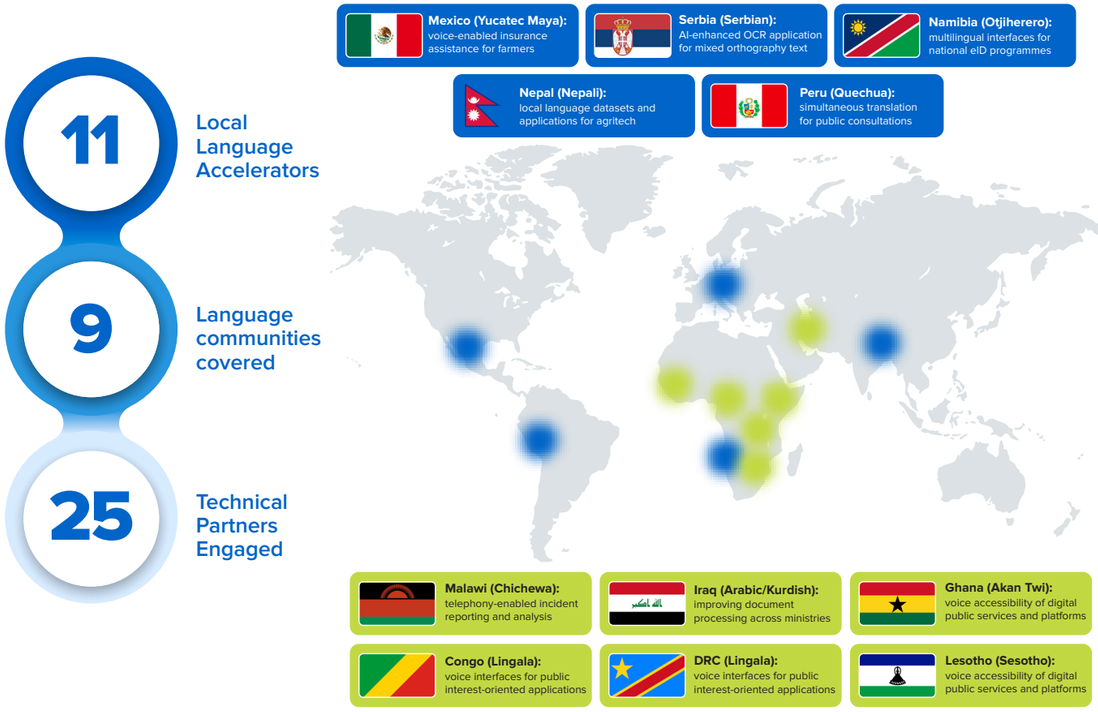
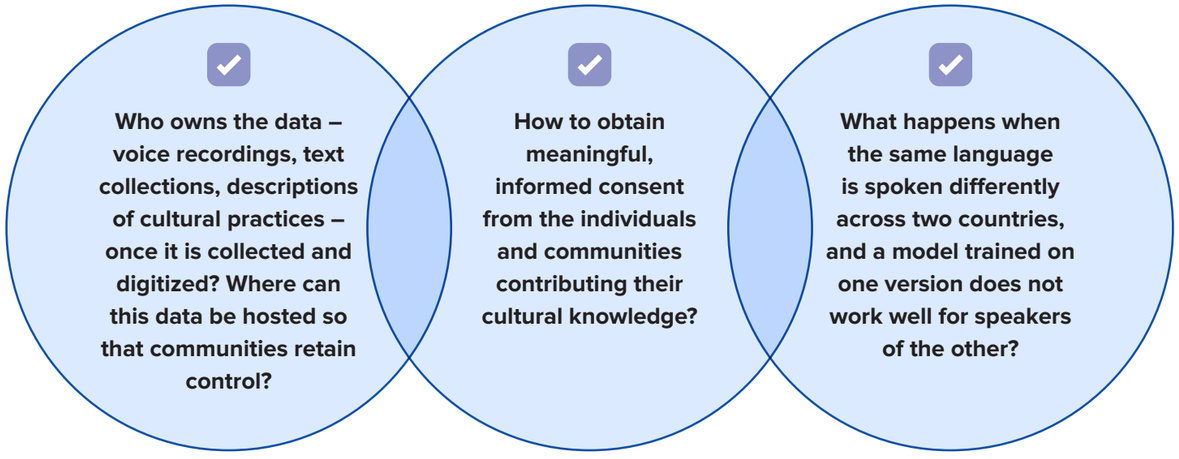


Figure 1: Current rollout of UNDP Local Language Accelerators around the world.

The project teams had to answer questions that are critical to local AI projects in general, and even more critical for those involving local languages:



As deeply contextual projects such as these progress in parallel, familiar patterns are emerging across continents. Language data is collected, communities contribute their expertise, and promising initiatives begin to take shape. Yet progress repeatedly stalls before these efforts become sustainable initiatives that can benefit people. Funding often covers data collection, but not the compute infrastructure required for model training and evaluation. Many initiatives lack a clear pathway into broader deployment.

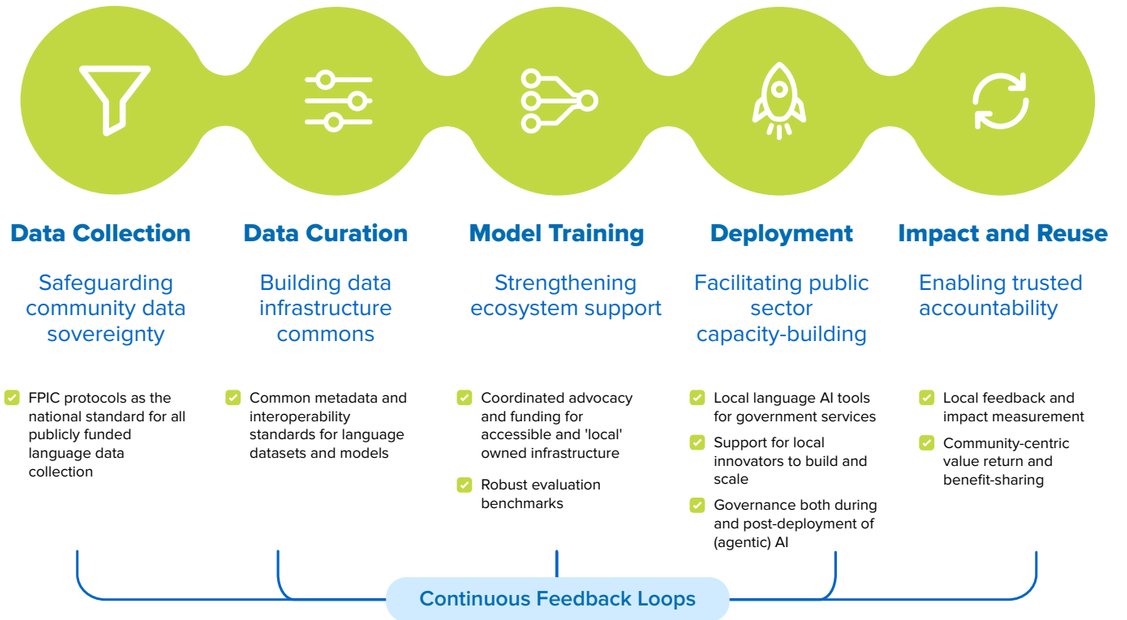


Figure 2: The Data-to-AI Value Chain

All of this demonstrates that building inclusive AI is not a simple pipeline where data flows in and AI tools flow out, but a process comprising five interdependent stages that form the **Data-to-AI Value Chain**.

The AI value chain in practice: **Current AI is building a cultural preservation pipeline**

Most AI progress on diversity has focused on language: adding more languages, localizing tools, and expanding coverage. But language is only part of the equation. Culture, its idioms, its knowledge-governance frameworks, is what gives language meaning. Cultural heritage, both tangible (objects, collections, monuments, sites) and intangible (oral traditions, rituals, craftsmanship, belief systems), matters just as much, and it is almost entirely missing from how AI is trained, governed, and deployed. A wealth of knowledge still lives in manuscripts, local media, and living memory, and very little of it is feeding into AI innovation today.

This gap matters for two reasons. AI development tends to advantage languages and cultures with lots of existing digitized data, which can leave the rest of the world's cultures underrepresented or erased. But the reverse is also true: AI has real potential to support cultural preservation, access, and transmission, helping safeguard cultures against threats like climate change, conflict, and decay, and carrying the wisdom of elders forward. Culture is not a static archive to protect from technology. It is a living system whose transmission AI can interrupt or support.

Current AI developed its Cultural Preservation Pipeline to guide innovation at the intersection of AI and cultural preservation, including linguistic diversity. The Cultural Preservation Pipeline is a framework that shares underlying logic with UNDP's Data-to-AI Value Chain, and the two can be read alongside each other. It helps innovators build responsibly at that intersection, keeping the interests of the communities whose knowledge and heritage are at stake front and centre. The Pipeline begins with data governance and community custodianship, the foundation for using AI to support inclusion and cultural revitalization. For culture-specific data projects, the two stages that follow are crucial: capturing cultural nuance accurately in digital form, and labelling, processing, and documenting it well. Getting this right can determine whether AI becomes supportive of a culture or disruptive to it.

Community custodianship is not a one-time checkpoint at the start of the process. Community evaluation and stewardship need to continue throughout: they keep systems auditable and correctable over time. Digitized heritage without community custodianship risks becoming extraction. Cultural datasets without local models and applications leave communities documented but not served. Strengthening the whole Pipeline, especially its weakest links, and the learning and cross-pollination between communities, is core to Current AI's public interest AI mission.

Getting this right within a single community is not sufficient; it is necessary to connect efforts that would otherwise stay isolated. Cultural preservation work tends to develop within a single language, geography, or institution, and those working on it may not be connected to others wrestling with similar problems elsewhere: how to digitize oral testimony without flattening it, how to govern access to sacred knowledge, how to build models that work with very little data. When cross-pollination is deliberate rather than accidental, it lets methods, frameworks, and infrastructure travel between communities. Accelerating this kind of connection is core to what Current AI invests in: building a global ecosystem for AI and cultural preservation.

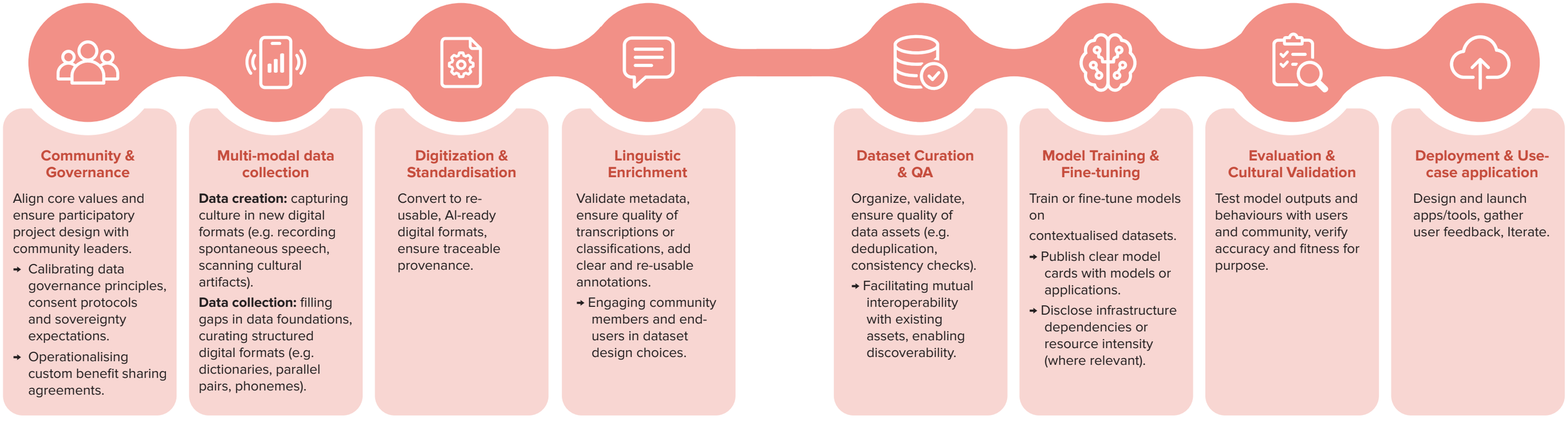


Figure 3: Charting Current AI's Cultural Preservation Pipeline

A GLOBAL SHIFT IN MINDSET: **TOWARDS SUSTAINABLE AI ECOSYSTEMS**

Linguistically and culturally inclusive AI might be stalling because it still gets treated as a technical or governance problem, when it is really an ecosystem-wide issue. Approaches that only centre technology tend to optimize some steps over others. They collect the data and train the model before building a product to release. **Ecosystem thinking asks what needs to be in place around that step for it to continue working well, now and in the future**. This tracks a broader critique in the AI literature: treating data work as a one-off step, rather than an ongoing responsibility is what causes 'data cascades' – failures that compound quietly and surface later as broken models.

This becomes most apparent when a project ends. Tech-first thinking leaves behind a dataset with nobody to carry it forward; ecosystem thinking leaves behind something that takes on a life of its own and enriches future projects: an institution with a long-term mandate, a graduate programme with its first class of students, a startup with its first government contract, a repository the next team can build on. **The five stages of the Data-to-AI Value Chain work the same way – each one only holds up if the others are in place**. A break at any point in the chain could prevent realization of value; simply fixing issues within one stage does not automatically address issues in others.

Table 1: Break points across the five-stage Data-to-AI Value Chain, and their impact

STAGE 1			
Process	**Where the process can fail**	**Why it matters**	**Strategic opportunity**
Data collection	Free, Prior and Informed Consent (FPIC) not consistently followed. Communities excluded from decisions about how their data is used. Women, elderly, and dialect speakers systematically underrepresented. Contribution ends with the grant.	AI systems are unusable due to poor representation of linguistic and cultural realities; extractive data practices reduce trust and long-term feasibility.	Investing in community-led data governance, inclusive participation, and long-term stewardship models ensures that communities retain meaningful rights over their contributed data.

Process	**Where the process can fail**	**Why it matters**	**Strategic opportunity**
	It is worth noting that data is not a single category. Voice recordings carry the strongest personal connection to the individuals who contributed them and require a rigorous consent framework. Published text may have institutional copyright but limited community governance claims. Lexical data raises deeper questions about who 'owns' the words of a language. These distinctions matter for how governance frameworks are designed at each stage.		Community-centric licensing frameworks such as NOODL[xiv] or Esethu[xv] determine how data is used downstream and ensure communities benefit from the AI systems they help to build.

STAGE 2			
Process	**Where the process can fail**	**Why it matters**	**Strategic opportunity**
Data curation	Annotation done by external actors without community review. Additionally, annotators performing essential labelling, transcription, and validation work (predominantly in the Global South[xvi]) are frequently underpaid, exposed to psychologically harmful content (without adequate support), and rendered invisible to supply chains of the institutions that benefit[xvii]	Poor-quality or incomplete datasets reduce model performance and reinforce biases, compounding the disadvantages of already under-resourced languages.	Building local annotation capacity, developing new data provenance standards and platforms, and high-quality natively co-authored language corpora.

Process	Where the process can fail	Why it matters	Strategic opportunity
	most from their labour. Provenance disappears once data enters a corpus. Datasets being translated from English rather than natively authored. But even then, it is important to note that the provenance and governance requirements differ depending on whether the data type is voice, text, or lexical, since each requires a different curation and licensing approach.	As agentic AI becomes more common, deployment governance must also address how agents are authorized for ongoing access to community data sources and how agent-mediated data use is audited, since these interactions are not covered by consent frameworks designed around static model training.	Also, data procurement, at the minimum, should require fair pay, safe working conditions, meaningful contracts, and worker representation for annotators.

STAGE 3

Process	Where the process can fail	Why it matters	Strategic opportunity
Model training	Compute funding limited to credits, not local infrastructure. Skills not built in-country. Every programme duplicates the same data pipelines. Community data trains commercial models with no value returned.	Countries remain dependent on external providers, and the economic value generated by community data flows outward rather than back to the communities that produced it.	Investing in shared compute infrastructure through regional partnerships, national compute nodes, and multilateral mechanisms.

STAGE 4

Process	Where the process can fail	Why it matters	Strategic opportunity
Deployment	Grants end before products reach people. Only limited, non-operational public spending on local language capabilities for public services. Prototype-to-production gap consistently underestimated. Low adoption where communities were excluded from co-design. As agentic AI grows, deployment governance must address how agents access community data on an ongoing basis and how that access is audited, since existing consent frameworks were not designed for this.	Public investments generate pilot projects or prototypes rather than long-term services that benefit communities.	Securing country deployment mandates and operational budgets that outlast the project funding cycle. Governance must extend post deployment to cover how AI systems are monitored, corrected, and held accountable over time, including who audits performance, how communities could contest outcomes, and what mechanisms ensure value continues to return to contributing communities. Additionally, local language capability in procurement as a pre-requisite standard is a direct way to make the above happen xviii.

STAGE 5			
Process	**Where the process can fail**	**Why it matters**	**Strategic opportunity**
Impact and reuse	No value-return mechanisms or revenue-sharing. No community satisfaction measurement with data collection processes or end-products. Impact tracking receives almost no dedicated funding. Speakers who contributed their data see nothing in return.	Benefits cannot be demonstrated, sustained, or scaled, and the loop between community contribution and community benefit never closes.	Establishing benefit-sharing frameworks that return value to contributing communities. Publishing annual progress reports that are accessible to communities and commissioning independent reviews and audits.

Taken together, these challenges point to a broader structural problem that extends beyond data availability or technical capability. Across 26 countries, UNDP's own AI Landscape Assessments repeatedly show that political ambition and technical talent are insufficient without robust governance and financing structures to connect and sustain them. Ecosystem insight-sharing is also required to minimize duplication and ensure the next team does not have to start from zero. Without these pillars, local innovation will struggle no matter how many strategies are published.

SPOTLIGHT ON DIFFUSION PATHWAYS: **VOICE DATA AND HYPERLOCAL DICTIONARIES IN THE DEMOCRATIC REPUBLIC OF THE CONGO (DRC)**

In the Democratic Republic of the Congo, UNDP's work on Lingala AI began with the #Elikya (which means hope in Lingala) initiative. The initiative set out to integrate Lingala into AI systems while preserving its cultural weight as a lingua franca spoken by more than 20 million people across the Democratic Republic of the Congo and more than 40 million across Central Africa. Funded through the AI Language Accelerator with support from the French government, #Elikya mobilized universities, research centre, technology hubs, and young innovators from across the Democratic Republic of the Congo to standardize linguistic data and develop locally adapted AI prototypes, including chatbots, translators, and voice assistants in Lingala. Field operations are currently active, with teams deployed in Mbandaka to collect voice data directly from communities. Equipment for recording and networking has been procured and is being staged for deployment, although connectivity in the field remains an ongoing operational constraint.

The primary data collection instrument is Voice-a-thon, a community-driven voice data platform developed by Fikiri, a Kinshasa-based innovation ecosystem, as part of UNDP's Local Language Accelerator Programme. Unlike crowdsourcing tools such as Mozilla Common Voice, Voice-a-thon is designed with the country-specific context in mind: it is built and maintained locally and gives communities a greater degree of control over how voice data is collected and validated. In addition, the project plans to integrate the Nwulite Obodo Open Data License (NOODL), a community-centred data licensing framework into this pipeline, with the aim of returning meaningful autonomy to the communities whose voices form the foundation of the corpus. In practice, that would mean communities retaining the right to determine who can access their recordings, for what purposes, and under what conditions, including the ability to exclude any uses or commercial applications to which they have not consented. It would also mean that the data does not simply leave the community and become an asset owned by a distant institution. To date, the platform has produced a linguist-validated corpus of one million parallel sentences in Lingala and French, providing the textual foundation for high-quality voice recordings.

These efforts reflect a broader model of AI diffusion grounded in a voice-first approach, which is not simply a technical choice but a commitment to inclusivity. In a context where literacy rates vary significantly across communities and where Lingala is primarily spoken rather than written, voice interfaces lower the barrier to interacting with digital services. A farmer in Mbandaka, a health worker in a remote clinic, or a community leader can engage with a voice-enabled service in their own language without needing to read or type. This is what it means to build AI for everyday settings: not tools designed for the digitally fluent, but systems shaped around how people communicate. A high-quality, community-sourced Lingala voice corpus is not only a technical asset. It is the prerequisite for use cases that matter: voice-enabled public services, health information lines, agricultural advisory tools, civic information systems, that would otherwise remain out of reach for millions of Lingala speakers. Progress in the Democratic Republic of the Congo is being built on reusable assets including a validated text corpus, a locally controlled data platform, community mobilization experience, and a planned equitable data licensing framework, all of which reduce the cost and risk of future interventions and can be replicated across sectors and regions.

LEVERAGING THE DATA TO AI VALUE CHAIN FOR SYSTEMATIC CHANGE: **A GOVERNMENT'S VIEW**

Across the globe, there is demand for linguistically and culturally inclusive AI systems. National AI strategies are frequently ambitious in this domain; they signal commitment to local AI development, open datasets, and local-language-enabled public services. Yet implementation takes time, and it is not always obvious where to start. Data collection and curation require coordinated mandates across government, spanning the ministries responsible for interior affairs, culture, education, and information and communications technology (ICT). The channels needed to coordinate across these institutions may not yet exist. Public records, broadcast archives, and education materials lack a release roadmap and a named owner, leaving them locked even where a strategy commits to openness. Across the system, unfit procurement practices privilege large external vendors over local innovators, startups, or research labs actively building public interest-oriented prototypes. A lack of clear and actionable accountability mechanisms make progress difficult to track. **Paired with the Data-to-AI Value Chain, Every Language Matters helps policymakers identify where their progress might be stalling and ensures that national ambition meets concrete collective action that nurtures local, linguistically and culturally inclusive AI**.

A strategy describes national intent and ambition. Yet its implementation requires something more specific: clearly defined actions, named owners, financing, partnerships, timelines, and accountability, preventing a 'stagnation gap', where a community stays a passive consumer of AI rather than a proactive builder of it.

The commitment framework for Every Language Matters bridges that gap, translating strategic ambition into practical commitments that governments can implement and monitor over time.

Governments already hold many of the levers of impact required to enable truly inclusive, locally empowering AI innovation. UNDP's five-step guidance to help governments make sense of what may be holding their local AI ecosystems back, and how to systematically re-align these incentives in favour of inclusive, open AI innovation.

1. **A cross-ministry governance mechanism with a named owner**. Evidence from UNDP's AI Landscape Assessments across 26 countries shows that mandate fragmentation is not a Language AI problem specifically; it is a common structural feature of AI rollout in government. Language AI mandates for instance may cut across ministries of culture, education, ICT, and communications. Without a named inter-ministerial owner and a mandate to coordinate, each ministry waits for direction that never comes. The governance mechanism does not need to be largely complex; it just needs to be specific: a named body, a defined scope, and the authority to make decisions that bind more than one ministry.

2. **A specific, time-bound public data release commitment with a transparency report**. Countries hold the largest language data repositories in their national territory: public records, broadcast archives, school curricula, civil registration documents, parliamentary debates. Data critically determines if, or how, implementation takes place. A commitment to audit those assets, publish a release roadmap, and release holdings under appropriate open licenses – with a public transparency report on what was released and why – transforms the resource landscape for a language and does not require new data collection. It also signals to the domestic AI sector that the government is a serious actor in the ecosystem.

3. **Procurement standards that require local language capability and independent evaluation benchmarks**. Without a procurement signal, there is no commercial imperative for vendors to invest in local language capability. Updating procurement standards to require multilingual capability declarations and independent evaluation benchmarks changes the market logic in various stages of the value chain. UNDP's Small States, Big Signals report found that procurement consistently emerged as the most immediate governance lever available. It gives local developers a path to government contracts, creates accountability for vendor performance, and minimizes the practice of deploying AI tools that do not work (for example in terms of language or cultural context) for the people they are meant to serve. It is equally important that the benchmarks used in procurement are designed with community input. Unfortunately, most benchmarks for under-resourced languages are translations of frameworks originally developed in high-resource language contexts, embedding normative choices about which harms count and how performance is measured.

4. **A deployment pathway that connects prototype-stage tools to government service delivery with an operational budget**. Working prototypes now exist in most Local Language Accelerator countries. What they consistently lack (and not unique to the LLAs) is an institutional home, a ministry owner, and a budget line that survives the end of the project that built them. A deployment commitment names the use case, names the ministry, and has the potential to commit a recurring operational budget so that tools built by universities, startups, and civil society organizations can eventually become public services.

5. **A measurement framework with a published baseline and an independent review mechanism**. Without a published baseline, progress cannot be verified. Without an independent review mechanism, accountability is self-reported. A measurement framework that defines indicators for AI service uptake, language coverage, equity of access, and value return to communities, combined with a public progress report and an independent review mandate, closes the loop between what was committed and what was delivered.

SPOTLIGHT ON CULTURE: **CURRENT AI GRANTEES**

While linguistic diversity in AI has begun to gain traction, the oral traditions, knowledge systems, and lived experience that give language its meaning remain largely absent from how AI systems are trained, governed, and deployed. Much of this knowledge was never digitized. It exists instead in manuscripts, local media, and living memory, invisible to AI by default.

Current AI's inaugural grant cohort was designed to help close this gap. Four organizations, spanning sub-Saharan Africa, the Middle East and North Africa, and the Brazilian Amazon, tackle cultural preservation from different starting points: the conditions that make ethical data collection possible; the infrastructure that makes cultural knowledge legible to AI systems; the sovereignty frameworks that keep communities in control throughout; and the tools to hold it all to account. The cohort was also designed for exchange between grantees, not only for each to work in parallel. What each produces – validated datasets, governance frameworks, mobilization methods, audit tools – can be picked up and used again elsewhere. Current AI encourages releasing what can be shared openly, whether that's the full output or specific components, so other organizations facing similar problems do not have to start from zero. This reflects how Current AI works as a grant-maker: creating shared infrastructure, not just funding individual projects, and encouraging open source and open sharing to move the field forward together. In the Brazilian Amazon and Cerrado, where Indigenous language data has often been scraped into AI training pipelines without consent, **Portal sem Porteiras** works with Indigenous communities to develop offline-first, locally hosted AI tools. Communities decide which tools get built, retain the data within the territory rather than on distant servers, and hold decision-making authority at every stage, from what gets recorded to how any resulting model is trained and deployed.

Arab cultural knowledge is largely missing from AI systems. The **Institute for Worldmaking**, based in Lebanon but working across the Middle East and North Africa is working to close this gap: assembling machine-readable, consented databases of Arab cultural knowledge, an auditable research agent, and governance frameworks that let Arab communities and diaspora researchers determine how their knowledge enters AI systems.

Working across the African continent, where language communities have historically had no visibility into AI systems, **Masakhane** brings them together to build datasets across more than 50 languages, extending into domain-specific data for health, farming, and education. The communities contributing to the data also shape how it gets structured and used, testing similar custodianship questions to other organizations elsewhere in the world, like Portal sem Porteiras.

Across the African continent, AI systems increasingly shape decisions like loan approvals, identity verification, and border crossings, often with little recourse for affected communities to contest outcomes. The African Internet Rights Alliance (AIRA) builds audit tools, designed for reuse by regulators, civil society, and policymakers, that verify whether the commitments made at deployment are being honoured. Whereas the other three organizations focus on how cultural knowledge and language enter AI systems responsibly, AIRA verifies whether those commitments hold up.

INTRODUCING EVERY LANGUAGE MATTERS: **A GLOBAL INITIATIVE TO BUILD A SHARED VISION OF INCLUSIVE LANGUAGE AI**

Building inclusive Language AI requires more than better datasets or better models. It requires an ecosystem approach that strengthens every stage of the Data-to-AI Value Chain, from data collection and governance through to deployment, impact, and reuse. **Every Language Matters** (ELM) is a global initiative designed to support that ecosystem. Every Language Matters (ELM) is a global initiative that organises that coordination around three main things: government commitments that are public, named, and backed by concrete plans; partner alignment of financing, infrastructure, and technical capacity around those commitments; and implementation partner support to help governments design, resource, build capacity in order to deliver on what they have committed to. UNDP's role is distinctive: with a presence in over 170 countries, long-standing relationships with governments at the policy and implementation level, and direct experience running language AI programmes on the ground, it is positioned to convene the actors, broker the commitments, and establish the accountability mechanisms that no single funder, researcher, or technology partner can provide alone.

The design of Every Language Matters is intentionally modular. Countries can engage where they are today, according to their priorities and level of readiness, while contributing to a shared vision of AI that leaves no one behind. Practically, governments make one or more commitments across the five commitment areas, name an owner, and work towards publishing a baseline (if none exists). Thereafter, rather than prescribing a single pathway, ELM enables governments and partners to better understand their own ecosystems, co-design impactful programmes, mobilize new investment, forge partnerships, and strengthen the full Data-to-AI Value Chain in ways that reflect a structured coordination mechanism that aligns with national contexts and priorities.

The role of Every Language Matters is not to replace national AI strategies or existing implementation programmes. Rather, it provides a shared framework through which governments and partners can organize collaboration, identify implementation priorities, mobilize investment, and track progress across the Data-to-AI Value Chain. **Commitment areas within ELM translate this ecosystem approach into practical actions that countries can adapt to their own context**. The framework organizes country commitments into five core pillars, each of which addresses a specific set of value chain failures. The table below shows what each area requires and what governments gain from it.

Table 2: Five commitment areas for governments within the Every Language Matters ecosystem approach

# 1 Commitment	What it addresses	What it requires	What governments gain
Data and Community Sovereignty *Stages 1 and 2*	Free, Prior and Informed Consent (FPIC) not consistently followed. Communities have no governance rights over their data. Public sector holds vast language assets that remain locked. Procurement does not require community benefit. Most available benchmarks embed normative choices about which harms count and how performance is measured that were made outside the community.	Adopt FPIC as the national standard. Establish community data cooperatives with legal recognition and benefit-sharing. Audit and release government-held language data. Reform procurement to require multilingual benchmarks. Ensure those benchmarks are developed with community input rather than translated from frameworks designed in high-resource language contexts.	Locked government data becomes a national AI asset. Procurement reform creates a domestic market for local language AI products. Communities become governance partners, not just data sources.

# 2 Commitment	What it addresses	What it requires	What governments gain
Reliable, locally governed AI infrastructure and model commons *Stages 2 and 3*	Every programme rebuilds the same pipelines. Compute is expensive and mediated through external actors. No shared repositories or evaluation standards. Knowledge built in one country is invisible to teams in another.	Designate a national node in an open model repository. Adopt common metadata and interoperability standards. Allocate compute for fine-tuning national languages. Mandate open evaluation frameworks.	Ministries draw from shared infrastructure rather than commissioning equivalent models separately. Costs fall. Locally owned AI infrastructure compounds in value across agencies and programmes.

# 3 Commitment	What it addresses	What it requires	What governments gain
Ecosystem and capacity-building *All stages*	Technical expertise is thin and concentrated among individuals. When a key researcher leaves,	Fund graduate programmes in local Language AI. Build peer mentorship networks connecting practitioners across countries.	A domestic technical workforce that governments do not have to outsource.

Commitment	What it addresses	What it requires	What governments gain
	the knowledge leaves with them. No graduate pipeline. No revenue model for local startups without donor funding.	Create startup revenue models through procurement set-asides and challenge funds. Invest appropriately in speech and text according to community context and use-case.	A local AI sector with a genuine path to market. Knowledge that stays in-country when researchers move on.

#4 Commitment	What it addresses	What it requires	What governments gain
Public service deployment *Stages 3 to 5*	Grants end before products reach people and communities. No institutional home for prototypes after the project closes. Communities excluded from co-design. Local innovators bypassed by large bilateral deals.	Commission one to two named use cases with a ministry owner and an operational budget. Publish open APIs and language datasets. Create procurement set-asides for local innovators. Co-design with communities from the start.	AI services that reach the full population including rural, elderly and non-literate speakers. A domestic AI economy rather than grant dependency. Visible service delivery improvements in health, agriculture, education and civic services.

#5 Commitment	What it addresses	What it requires	What governments gain
Measurement and Accountability *Stage 5*	No measurement framework. No published baseline. Impact tracking has almost zero dedicated funding. Communities who contributed data see no return. The loop never closes.	Publish a KPI framework covering AI uptake, language coverage, equity and value return. Establish baselines. Annual public progress reports including benefit-sharing data. Fund independent review.	A credible, published record of progress that attracts further investment. Accountability to communities that builds trust. Progress data that positions the country as a leader in the inclusive AI conversation globally.

The five commitment areas above focus on what governments must build on the demand-side of the Language AI ecosystem. There is a parallel set of obligations on the supply-side where ELM's design creates market signals and governance requirements for frontier model developers, cloud compute providers, and platform owners to leverage – one such example being procurement.

LINGUISTIC AND **CULTURAL DIVERSITY AS THE NEW AI FRONTIER**

The initiative is asking countries for commitments that unlock value that governments already possess. The most important of these is data. Several governments in linguistically diverse countries already hold the largest language data repositories for their own context; for example, broadcast archives, school curricula, or civil registration documents. Much of this data is exactly what domestic AI development needs, but it usually has not been unlocked or enhanced for AI in a structured way. **Releasing this kind of archive under appropriate licenses and personal data protections and other legal provisions, with clear public documentation of what is being released and why, has the potential to create an immediate resource for local AI innovation**.

What governments gain by joining Every Language Matters:

Beyond data, the initiative offers four opportunity areas that matter to governments and other local AI ecosystem actors:

1. **Improved institutional coherence and coordination**: an inter-ministerial owner and a shared framework ends the waiting game between ministries. Language AI policy stops being everyone's responsibility and institutionalization is centred.

2. **A robust market signal to spur local innovation:** procurement reform that requires multilingual capability sends a clear signal to the local AI sector that there is a government demand for language products. That signal alone changes what local startups choose to build – preferring tools that plug into the relevant services and improve service delivery.

3. **An international spotlight:** participating governments are brought into international forums on inclusive AI such as during the high-level week of the UN General Assembly. For many governments, this is an important immediate and visible return on their commitment so that such work gains prominence.

4. **New means of making public service delivery more inclusive and efficient:** AI tools built only for higher-resource languages (and to continue investment only in similar language data pipelines) leave millions of people without access to health advice, legal information, agricultural guidance, and civic services. Research and investments in solutions (and infrastructure) that surface under-resourced languages changes that, and governments have a key role to play.

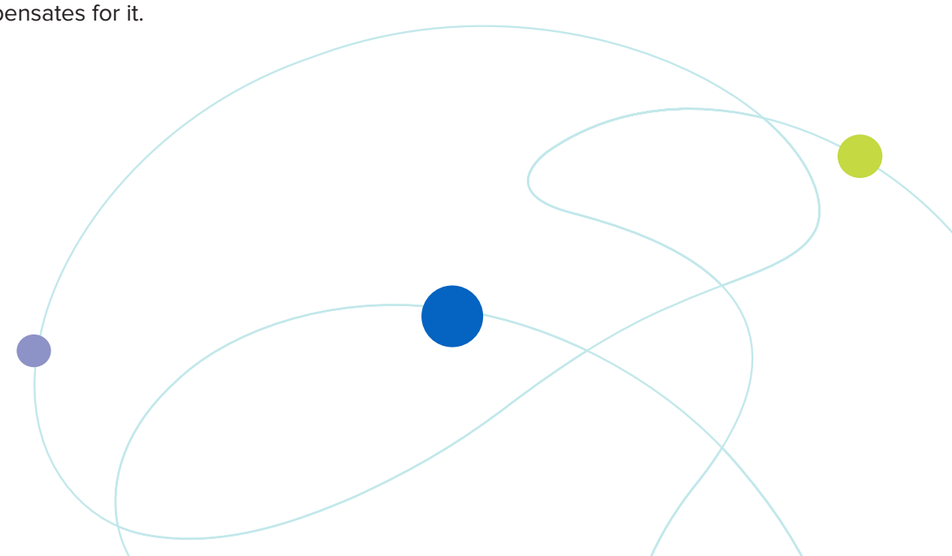
CONCLUSION: **LET'S BUILD PUBLIC, INCLUSIVE AI TOGETHER**

Language is the interface through which people communicate, reach public services, exercise rights, learn, and participate in economic life; culture is the glue that gives it meaning. When AI systems reflect only a handful of languages and cultures, they embed exclusion into the infrastructure that will mediate the next decades of public life. Linguistically and culturally inclusive AI is not a matter of coverage or fairness in abstract; it determines whether the benefits of digital transformation reach most of the world at all.

The Data-to-AI Value Chain explains why that inclusion so often stalls between a promising pilot and sustained public impact. Success depends not only on collecting high-quality language data, but on the governance, infrastructure, institutions, financing, and partnerships that connect every stage from collection to deployment, and the cross-pollination that can happen between innovators and institutions. Where those connections are missing, the chain breaks – and no amount of data, talent or compute compensates for it.

Two years of Local Language Accelerator implementation across eleven countries point to the same conclusion: the Language AI value chain is broken not for lack of ambition, talent, or data, but for lack of the mechanisms needed to bring these elements together. Every Language Matters is a global initiative that addresses this by helping policymakers convert ambition and strategy into concrete commitments across five areas: data and community sovereignty, AI infrastructure and model commons, ecosystem and capacity-building, public service deployment, and measurement and accountability.

Governments, researchers, funders, technology partners, and communities each hold part of the solution, and Every Language Matters exists to turn these individual efforts into a productive, sustainable ecosystem of trust and innovation at a global scale that promotes discoverability of promising work, mobilizes investment, and inspires new partnerships.



We invite governments, funders, and technology partners to join the initiative and make a commitment in at least one area this year, so that current strategies and research finally reach the communities who contributed their voices, stories, and time – and who are still waiting.

Read about how to join **Every Language Matters**:

https://www.undp.org/digital-innovation/every-language-matters

REFERENCES

African Academy of Languages. (n.d.). *Public Inventory of African Languages and Resources* [Data set]. African Union Commission. https://docs.google.com/spreadsheets/d/1fq2zb9NFBJdqg15hMckUQa1drYEauiAWfZXgNwOVGE0/

Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D. R., Smith, N. A., & Tsvetkov, Y. (2023). Do all languages cost the same? Tokenization in the era of commercial language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9904–9923). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.614

AI Singapore. (2023). *Southeast Asian Languages in One Network (SEA-LION)*. https://sea-lion.ai/

Alabi, J. O., Hedderich, M. A., Adelani, D. I., & Klakow, D. (2025). Charting the landscape of African NLP: Mapping progress and shaping the road ahead. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 27807–27841.

Azin, M. H., & Zandhessami, H. (2025). Strategic alignment patterns in national AI policies. *arXiv*. https://arxiv.org/abs/2507.05400

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, *19*(1), 43. https://doi.org/10.5334/dsj-2020-043

Doumbouya, M., Einstein, L., & Piech, C. (2021). Using radio archives for low-resource speech recognition: Towards an intelligent virtual assistant for illiterate users. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 14757–14765. https://ojs.aaai.org/index.php/AAAI/article/view/17733

Egyptian Ministry of Communications and Information Technology. (n.d.). *Global AI governance frameworks: A comparative study*. https://ai.gov.eg/SynchedFiles/en/Resources/Global%20AI%20Governance%20Frameworks%20A%20Comparative%20Study.pdf

Ekgren, A., Gyllensten, A. C., Stollenwerk, F., Öhman, J., Isbister, T., Gogoulou, E., Carlsson, F., Casademont, J., & Sahlgren, M. (2023). GPT-SW3: An autoregressive language model for the Scandinavian languages. *arXiv*. https://arxiv.org/abs/2305.12987

Etori, N. A., Dawson, M., & Gini, M. (2024). Double-edged sword: Navigating AI opportunities and the risk of digital colonization in Africa. In *MWAIS 2024 Proceedings* (Paper 25). AIS Electronic Library. https://aisel.aisnet.org/mwais2024/25/

Gulla, J. A., Liu, P., & Zhang, L. (2026). NorwAI's large language models: Technical report. *arXiv*. https://arxiv.org/abs/2601.03034 Infocomm Media Development Authority. (2023, 4 December). *SG to develop Southeast Asia's first LLM ecosystem* [Press release]. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/sg-to-develop-southeast-asias-first-llm-ecosystem

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Katende, R. (2026). Rethinking data-efficient artificial intelligence for low-resource settings. *Machine Learning with Applications*, *23*, Article 100796. https://doi.org/10.1016/j.mlwa.2025.100796

Kummervold, P., Wetjen, F., & de la Rosa, J. (2022). The Norwegian Colossal Corpus: A text corpus for training large Norwegian language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3852–3860). European Language Resources Association. https://aclanthology.org/2022.lrec-1.410/

Kyobe, M., & Opoku, R. (2025). "We know what we are doing": The politics and trends in artificial intelligence policies in Africa. *African Affairs*, *124*(495), 437–455. https://doi.org/10.1080/00083968.2025.2456619

Laiyk, N., Orel, D., Joshi, R., Goloburda, M., Wang, Y., Nakov, P., & Koto, F. (2025). Instruction tuning on public government and cultural data for low-resource language: A case study in Kazakh. *arXiv*. https://arxiv.org/pdf/2502.13647

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R., & Modupe, A. (2020). Investigating an approach for low resource language dataset creation, curation and classification: Setswana and Sepedi. *arXiv*. https://arxiv.org/abs/2003.04986

Monteiro, B., & Kumpf, B. (2023). *Innovation labs through the looking glass: Experiences across the globe*. Observatory of Public Sector Innovation. https://oecd-opsi.org/blog/innovation-labs-through-the-looking-glass/

Nigatu, H. H., Tonja, A. L., Rosman, B., Solorio, T., & Choudhury, M. (2024). The Zeno's paradox of 'low-resource' languages. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17753–17774). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.983

NTNU. (n.d.). *Large language models at University of Oslo*. https://www.ntnu.edu/norwai/large-language-models-at-university-of-oslo

Occhini, G., Tanaka-Ishii, K., Barford, A., Tikochinski, R., Hu, S., Reichart, R., Zhou, Y., Claus, H., Petti, U., Vulić, I., Debnath, R., & Korhonen, A. (2026). *Artificial intelligence is creating a new global linguistic hierarchy* (arXiv:2602.12018). arXiv. https://doi.org/10.48550/arXiv.2602.12018

OECD. (2025). Implementation challenges that hinder the strategic use of AI in government. In *Governing with artificial intelligence*. https://www.oecd.org/en/publications/governing-with-artificial-intelligence_795de142-en/full-report/implementation-challenges-that-hinder-the-strategic-use-of-ai-in-government_05cfe2bb.html

Opoku, R. (2026, 20 March). What Nigeria can learn from Ghana and South Africa as it drafts its AI strategy. *Techpoint Africa*. https://techpoint.africa/insight/ghana-south-africa-ai-strategy-lesson-for-nigeria/

Research Data Alliance International Indigenous Data Sovereignty Interest Group. (2019). *CARE Principles for Indigenous Data Governance*. The Global Indigenous Data Alliance. https://www.rd-alliance.org/wp-content/uploads/2024/03/CARE20Principles20for20Indigenous20Data20Governance_OnePagers_FINAL20Sept2006202019.pdf

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445518

Thomson Reuters Institute. (2026, 23 February). *New data reveals AI governance gap between policy and practice, creating ESG risks*. https://www.thomsonreuters.com/en-us/posts/sustainability/ai-governance-gap-esg-risks/

United Nations Development Programme. (2026). *Reading AI readiness backwards: What country experience reveals once AI adoption is already underway*. AI Hub for Sustainable Development. https://www.undp.org/publications/reading-ai-readiness-backwards-country-insights-ai-adoption-and-implementation

United Nations Development Programme. (2026). *Small states, big signals: What AI adoption in practice reveals about trust, safety, and AI performance globally*. United Nations Development Programme AI Trust & Safety Re-Imagination Programme. https://www.undp.org/publications/small-states-big-signals-what-adoption-practice-reveals-about-trust-safety-and-ai-performance-globally

Yew, R.-J., Creasey, K. E., Curtis, T. L., & Venkatasubramanian, S. (2026). The commodification of AI sovereignty: Lessons from the fight for sovereign oil. *arXiv*. https://arxiv.org/abs/2601.11763


UN
DP


Current AI