Tonative | DEEP LEARNING INDABA

WORKSHOP GUIDE

DEEP LEARNING INDABA 2026 · NIGERIA

Building Africa's Human Data Infrastructure

Capacity building for community-led dataset creation and sustainable AI pipelines.

LOCATION	DURATION	FORMAT	HOSTED BY
DLI 2026, Lagos, Nigeria	90 minutes	Forums & dialogues	Tonative Africa

ABOUT THE WORKSHOP

A working session on the people, skills, and systems behind African AI data

The rapid growth of generative AI has intensified demand for high-quality datasets, yet progress in African AI remains constrained by gaps in **human data infrastructure.** This includes the people, skills, and coordinated pipelines required to create, translate, validate, and maintain datasets for African languages and contexts at a large scale.

This workshop is hosted by **Tonative Africa** at the **Deep Learning Indaba 2026** in Lagos, Nigeria, and it focuses on capacity building for creators, translators, validators, and annotators. Through a short talk, breakout discussions, and collaborative roadmap design, we will explore best practices for dataset creation, quality assurance, and long-term capacity development across African language communities.

The session aims to produce shared guidelines, identify priority challenges, and co-develop a roadmap for strengthening Africa's sovereign, sustainable, and locally owned AI ecosystems.

HOW TO USE THIS GUIDE

You'll join one of three breakout groups. This guide walks you through the full session and previews the discussion questions for each theme. Skim the session you will be attending before it begins.

WHO WE ARE

Tonative - African context intelligence

Tonative trains African data curators, builds custom train & eval datasets, and ships AI tools that improves AI performance on African languages and local contexts. We work across **20+ languages and 5 industry domains** with a growing network of curators.

TONATIVE DATA ACADEMY

A 6-week program for African Language Data Curators

The course covers seven modules, which includes: Introduction to Language and Linguistics; Forms of NLP Datasets for Training, Tuning and Evaluating AI Models; Audio and Text Data Curation; Data Validation and Quality Checks; Dataset Documentation; Ethics and Governance for African datasets, and a hands on data curation lab session. Graduates earn a certificate and can join our team of curators.

TOOLS WE USE

The stack behind our pipelines

TonalKeys
AFRICAN LANGUAGE KEYBOARD

A free, lightweight Chrome extension for typing tone marks and diacritics across all African languages.

MTStudio
TRANSLATION ENGINE

Tonative's African-language translation engine. User upload a CSV and translate datasets into African languages to accelerate parallel-text creation.

Label Studio
SELF-HOSTED ANNOTATION

Label Studio is an open-source annotation platform by HumanSignal, that we self-host giving us full control over data, and quality validation.

SESSION AGENDA

Ninety minutes, five sections

01 Opening — 5 min

Welcome, session goals, and an introduction to Tonative and the Data Academy for African Languages.

02 Guest Talk — 10 min

A short invited talk by Dr. Lilian Wanzare on challenges in African dataset creation, community capacity building, and sustainable data pipelines.

03 Breakout Discussions — 30 min

Participants split into groups around three key themes across the data pipeline (detailed below).

04 Collaborative Roadmap Building — 30 min

Groups share key insights and co-develop a shared roadmap for strengthening capacity and pipelines.

05 Conclusion & Next Steps — 5 min

Key takeaways, opportunities for collaboration, and post-Indaba follow-up plans.

THE BREAKOUT SESSIONS

01 THEME
Capacity Building for Data Curation

How language speakers become skilled data curators and how curation can be a sustainable livelihood rather than disposable, gig-style tasks.

START HERE · DEFINE YOUR TERMS
Training · qualification · empowerment. What does each mean for a data curator?

DISCUSSION QUESTIONS

- How does a language speaker become a data curator? Is formal training required to get started?
- If training is required, what standardised topics or curriculum should it cover?
- How can we structure real career progression for curators, rather than treating annotation as disposable, gig-worker tasks?
- What would it take to make data annotation a sustainable, respected livelihood in your community?

02 THEME
Tools & Infrastructure

The software (annotation platforms, libraries, frameworks e.g. Label Studio, Hugging Face) and the underlying hardware, cloud, storage, and connectivity needed to run them reliably across diverse African contexts.

START HERE · DEFINE YOUR TERMS
Tools vs. infrastructure. What falls into each, and where do they overlap?

DISCUSSION QUESTIONS

- What tools do you rely on? Are they mostly open source or proprietary? For proprietary tools, are they custom or paid?
- How is tooling funded? By the community or through grants? Does cost dictate which tools you can use?
- Do these tools support local languages? If not, how do you work around it?

- What infrastructure bottlenecks do you face (e.g. high cloud-hosting and API costs, connectivity)?
- How can we pivot toward offline-first, self-hostable annotation and collection tools for remote or rural data creators?

03 THEME
Governance, Ethics & Policy

The rules and rights around data with a focus on licensing: who owns African-language datasets, how they can be reused, and what equitable frameworks look like.

START HERE · DEFINE YOUR TERMS

Data governance · ethics · policy, and an introduction to data licensing.

DISCUSSION QUESTIONS

- Have you created a dataset before? If so, what licensing did you use, and why?
- What do you understand about data licensing, and which licensing frameworks exist for language data?
- Commercial vs. open source: are international frameworks like Creative Commons practical for African datasets?
- What would a more equitable licensing framework for African datasets look like?
- How do annotators and communities navigate data-ethics inequities between different geographic regions?

SESSION 4 & 5 · BRINGING IT TOGETHER

From discussion to a shared roadmap

After the breakouts, every group shares its key insights and we co-develop a shared roadmap for strengthening capacity and building sustainable pipelines.

SCAN

Attendance Form

Register your attendance for the session.

SCAN

This Workshop Guide

Get the PDF and session materials.