

Gap Analysis of Data Playbooks for AfricaNLP

*A study of 86 playbooks, handbooks, and methodology guides
against the needs of AfricaNLP dataset builders*



Prepared by: HausaNLP, Bayero University, Kano, Nigeria
EthioNLP, Bahir Dar University (ICT4D), Bahir Dar, Ethiopia

Submitted to: Masakhane African Languages Hub

Grant: Annotation and Playbook Grant

Report date: 17 July 2026

Table of Contents

1	Executive Summary	3
2	Every Project Is a First Project	3
3	How We Mapped the Landscape	5
3.1	Scope and search	5
3.2	The 13 stages and how we scored them	5
4	The Landscape in Four Traditions	6
4.1	Documentation and governance frameworks	7
4.2	Generic process handbooks	7
4.3	The language documentation tradition	7
4.4	The African wave, including our own work	8
5	Reading the Gap Stage by Stage	9
5.1	Guidance stops at text and speech	11
6	The Playbook We Will Write	11
7	Conclusion	13
	Appendix A: Full Coverage Matrices	13
	Appendix B: Scoring Rules and Evidence References	23
	Appendix C: Abbreviations	24



1 Executive Summary

Building a high-quality dataset requires more than a good annotation tool. It requires a process: how to plan the project, engage the speaker community, recruit and pay annotators, obtain consent, collect speech and text, write annotation guidelines, control quality, and document and publish the result. Well-resourced languages inherit this process knowledge from decades of institutional practice. African-language teams, most of them building their first dataset, inherit almost none of it in usable form.

We reviewed 86 playbooks, handbooks, methodology papers, documentation frameworks, and governance instruments that a team could turn to for this guidance, including 34 resources created specifically for African languages, and scored each against 13 stages of the dataset lifecycle. The most complete resources in the audit are recent African annotation and benchmark methodology papers: AfriHate documents 11 of the 13 stages, more than any other resource reviewed, several others reach 9 or 10, and six of the eight resources worldwide that document fair annotator payment are African annotation projects. We know this line of work from the inside. Our own teams built NaijaSenti, AfriSenti, AfriHate, SemRel2024, BRIGHTER, EthioEmo, the Amharic hate speech corpora, and the Hausa Visual Genome, HaVQA, and Afri-MCQA vision and multimodal benchmarks, and they are scored here under the same rules as everything else.

That 11-of-14 ceiling is exactly the point. Even the most complete process documentation in the field reaches 11 of 13 stages, lives inside one-off project papers that were never written to be followed, and leaves speech collection, offline operations, and tool selection to other literatures that do not connect back. Nor is 11 the normal case: the median resource covers 4 of the 13 stages, so a data creator must today gather their process from many papers and manuals at once, from traditions that do not cite each other, because there is no single document to go to. Across the full landscape, 38 of the 86 resources cover 3 stages or fewer, only 16 give guidance on operating with unreliable connectivity, and only 11 are step-by-step playbooks of any kind, of which exactly one, the data.org Digitisation of Oral Data playbook produced with DSFSI, the University of Lagos, and Data Science Nigeria, addresses African language data directly. Guidance is also locked to two data types: beyond text and speech, the only full modality coverage anywhere comes from a handful of recent African vision-language benchmark papers, and OCR, video, and sign language are barely touched.

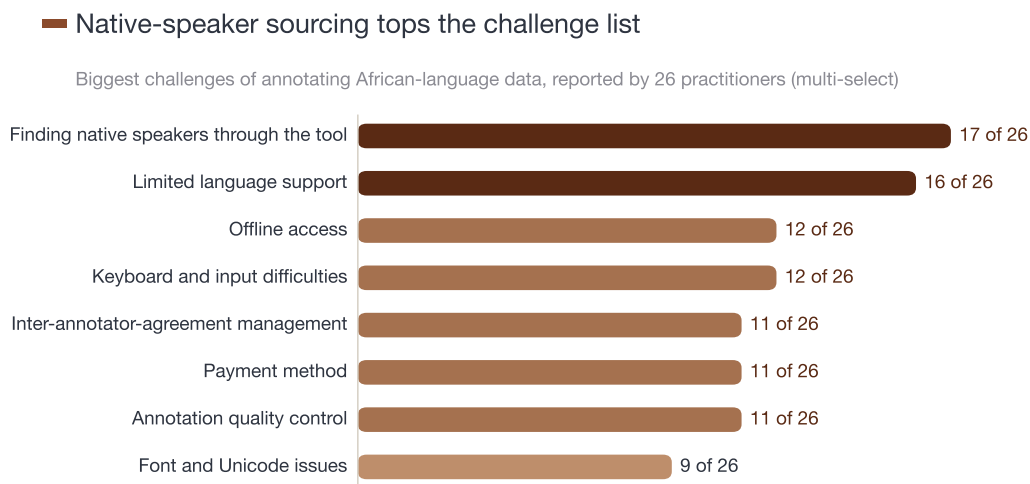
The answer we propose is the Masakhane Hub Data Playbook: a step-by-step, openly licensed guide to building African-language datasets, covering all 13 stages our study identifies and written as the process companion to AfriAnnotate, the annotation platform proposed under the same grant. The playbook will turn what our teams and our neighbours learned dataset by dataset, from recording practice in the fieldwork tradition to fair payment in our own annotation projects to community licensing in NOODL and Esethu, into one continuously maintained document that a first-time team can follow from idea to published dataset.

2 Every Project Is a First Project

African-language NLP (AfricaNLP) teams do not fail for lack of effort; they fail for lack of transmitted process. Most African languages have no existing datasets, so nearly every project

is a first project, run by researchers, students, and community organizations who must invent their workflow as they go.

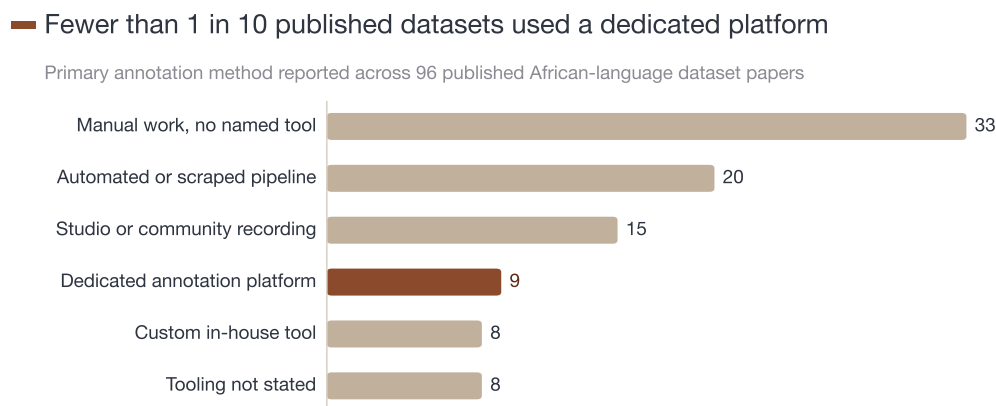
Two bodies of evidence, gathered for this grant's companion study of annotation tools, show what improvisation costs. First, a survey of 26 AfricaNLP practitioners asked for the hardest problems in annotating African-language data. The answers are process problems before they are tooling problems: finding native speakers came first (17 of 26), followed by limited language support (16), offline access (12), and keyboard and input difficulties (12), with a three-way tie among agreement management, payment method, and quality control (11 each).



Source: AfriAnnotate needs-assessment survey, July 2026 (n = 26).

Figure 1: Needs-assessment survey (n = 26; multi-select).

Second, an audit of 96 published African-language dataset papers recorded how datasets actually get built. Only 9 of 96 report using a dedicated annotation platform; 33 describe manual work with no named tool, and 8 do not state their tooling at all. Methodology gets the same treatment: the process is described after the fact, in the little space a paper allows, and almost never by pointing to a shared, citable guide that another team could follow. Each paper's lessons remain locked in that paper. We know this pattern intimately because our own dataset papers, among the best-documented in the field, are part of it.



Source: dataset-paper audit, July 2026. The six categories account for 93 of the 96 papers.

Figure 2: Dataset-paper audit. The highlighted bar is papers using any dedicated platform.

The consequence is a community that rebuilds its process from scratch for every dataset, repeats the same early mistakes, and publishes its lessons where the next team cannot find them. No existing survey maps the available guidance against these needs. This study does, and it motivates a purpose-built playbook to close the gap it finds.

3 How We Mapped the Landscape

3.1 Scope and search

We defined “playbook” broadly: anything a team could consult for process guidance. The audit includes self-described playbooks and handbooks, dataset documentation frameworks and templates, release checklists, funder guidance, language documentation manuals, methodology papers that function as reusable guides, and the African governance instruments that constrain or enable dataset work, from the African Union’s Continental AI Strategy and Data Policy Framework to national data protection laws and the community licensing frameworks NOODL and Esethu. African resources were prioritized in the search, which covered data.org, Masakhane, DSFSI, Lacuna Fund, AI4D Africa, Mozilla Common Voice, CLEAR Global, SADI-LaR, the African Union’s instruments, our own groups’ published methodologies, and the individual methodologies of major African dataset projects, alongside the global literature. The result is 86 resources: 34 African-specific, 15 partially African, and 37 generic.

3.2 The 13 stages and how we scored them

Each resource was scored against 13 stages of the dataset lifecycle, derived from the constraints in Section 2 and checked against the practitioner survey. Table 1 states each stage and its justification. A cell is scored covered (the resource gives substantive, actionable guidance), partial (the stage is touched but not useably treated), or not covered, always from the resource’s public documents at the report date. A covered score records that usable guidance exists, not that it targets African languages; African specificity is recorded separately. Separately from the 13 stages, we also recorded which data types each resource gives real guidance for: text, speech and audio, vision (images), documents and OCR, video and sign language, and cross-modal work that ties audio, video, and text together. The results are in Section 5.1 and Matrix 3.

Table 1: The 13 comparison criteria (lifecycle stages) and the justification for each.

Stage	What it means	Why it is a criterion
Project planning	Scoping, budgeting, and staging a dataset project in a low-resource setting	Most African-language projects are first projects; poor scoping locks in workflow debt for the corpus lifetime.
Community engagement	Participatory design with the speaker community, from goals to benefit sharing	African-language data comes from communities, and extractive collection is both unethical and unsustainable.
Recruiting native speakers	Finding, selecting, and onboarding native-speaker annotators and contributors	Respondents rank finding native speakers as their hardest problem (17 of 26).
Paying contributors	Fair compensation guidance, including cross-border payment mechanics	Payment is a top-tier survey challenge (11 of 26) and an ethical obligation.
Consent and ethics	Contributor consent, ethics review, and data protection through the project	Documented consent is a prerequisite for lawful and ethical dataset release.
Licensing and release	Choosing and applying a machine-readable license before release	A dataset released without a clear license is legally and operationally under-defined.
Data collection	Sourcing and collecting raw data in any modality: recording speech, sourcing text, gathering images, documents, and video, under license constraints	Every dataset starts with collection, each modality is its own craft, and Matrix 3 records which of them each resource covers.
Annotation guidelines	Writing task guidelines and training annotators on them	Guideline quality bounds dataset quality; most teams write their first guidelines unaided.
Quality control	Agreement measurement, adjudication, and review workflows	Agreement management and quality control tie among the top survey challenges (11 of 26 each).
Tooling guidance	Selecting and operating annotation and collection tools	Our companion study shows tool choice is consequential and poorly supported.
Scripts and language specifics	Orthography, scripts, keyboards, dialect and variety handling	Limited language support and input difficulties are top-ranked survey obstacles.
Offline and field operations	Running distributed teams on unreliable, expensive connectivity	Offline access ranks among the top survey challenges (12 of 26).
Documentation and publication	Datasheets, metadata, and publishing to archives and dataset hubs	Undocumented datasets are hard to trust, find, and reuse.

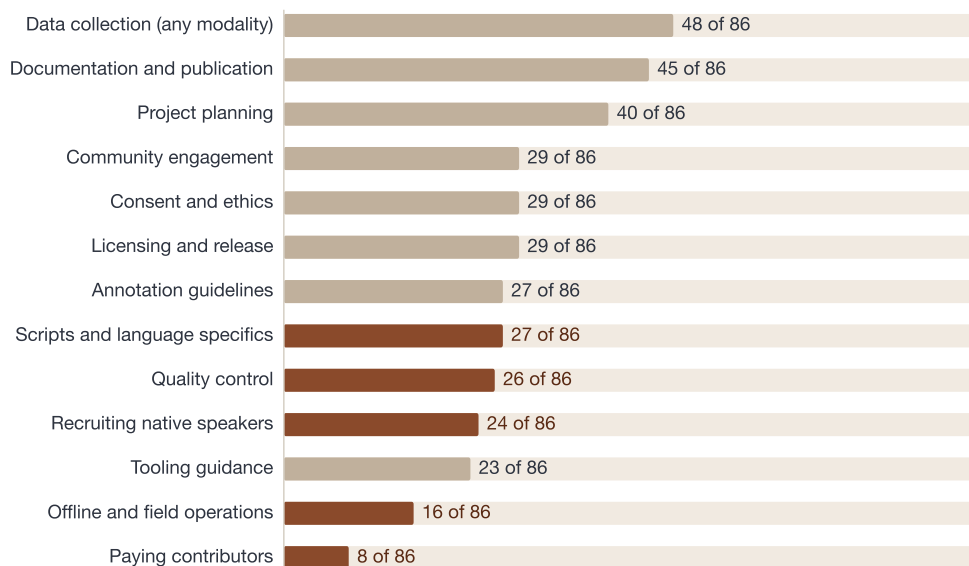
Note: We selected the criteria, and a different set could rank resources differently. We also declare an interest plainly: several of the audited resources are our own publications. They were scored under the same rules and from the same public documents as everything else, and where our own papers fall short of a stage, the matrices in Appendix A say so.

4 The Landscape in Four Traditions

The 86 resources are not a uniform genre. They cluster into four traditions, each strong in one place and silent in another, and the coverage pattern in Figure 3 follows from that clustering.

■ The stages practitioners struggle with most are the least covered

Number of the 86 reviewed resources fully covering each of the 13 stages;
terracotta bars mark the needs African-language teams rank highest



Source: playbook landscape audit of 86 resources, July 2026. Partial coverage is not counted; the median resource covers 4 of the 13 stages. Data collection counts any data type; Figure 8 breaks it down by type.

Figure 3: Coverage audit of 86 resources across the 14 lifecycle stages. Partial coverage is not counted.

4.1 Documentation and governance frameworks

Datasheets for Datasets, Data Statements, the Data Cards Playbook, Croissant, FAIR and CARE, the venue checklists, and the indigenous data sovereignty frameworks tell a team how to describe, license, and govern a dataset. They are the reason documentation and publication is the best-covered stage in the audit (45 of 86) and consent and ethics among the strongest (29). They are also, almost uniformly, silent on production: none teaches a team how to record speech, run a quality workflow, or find an annotator. They document the artifact, not the work.

4.2 Generic process handbooks

The Turing Way, the IFRC Data Playbook, the annotation methodology books (the MATTER cycle, the Handbook of Linguistic Annotation), Universal Dependencies' contribution guidelines, and the ELRA and LDC validation traditions teach process, and some of them teach annotation process well. But they assume the conditions of well-resourced institutions: review boards, reliable connectivity, payroll, and usually English. Cross-border payment, phone-first annotators, non-Latin keyboards, and data-cost budgeting do not appear. The process knowledge is real, but it was written for conditions African teams do not have.

4.3 The language documentation tradition

SIL's methodologies, the ELDP grantee guides, the fieldwork handbooks, and the CoLang materials carry decades of practice on exactly what the handbooks lack: recording under field

conditions, ethics with speaker communities, and working where connectivity fails. The ELDP guides cover 9 of 13 stages, among the broadest in the audit. But this tradition stops where NLP begins. Its annotation is transcription and glossing, its publication target is a language archive, and machine learning tasks, agreement metrics, and dataset hubs are out of scope. Only 8 of the 86 resources fully cover both annotation guidelines and speech recording (Matrix 3); the gap between the two traditions runs straight through the middle of an African speech project.

4.4 The African wave, including our own work

The youngest tradition is African-built, and it now holds the most relevant knowledge in the audit. It has three layers.

The first layer is project methodology. Recent African dataset projects document what almost nothing else in the landscape does: real annotator recruitment across dozens of languages, payment benchmarked above local wages, guideline design for code-switched and multi-script text, agreement measurement with honest numbers, and distributed teams run across countries. The audit's scores reflect it: AfriHate covers 11 of 13 stages; EthioEmo and the Amharic 2024 corpus 10 each; AfriSenti, Afri-MCQA, SemRel2024, and BRIGHTER 9 each, alongside Masakhane's participatory research framework at 9, NaijaVoices' reciprocal data farming, and the methodologies of Kencorpus, Makerere's East African lessons, the AI4D program, and the Mozilla East Africa community report close behind. Six of the eight resources worldwide that fully document annotator payment are African annotation projects. For transparency: many of these are our own. The HausaNLP and EthioNLP teams behind this report, working within Masakhane, produced NaijaSenti, AfriSenti, AfriHate, SemRel2024, BRIGHTER (Best Resource Paper at ACL 2025), EthioEmo, EthioLLM, the Amharic hate speech corpora, and the Hausa Visual Genome, HaVQA, and Afri-MCQA (Best Social Impact Paper at ACL 2026) vision and multimodal line, with ArtELingo-28 carrying our member Ibrahim Said Ahmad and further contributions to CVQA; all are scored under the same rules as every other resource.

The second layer is guidance proper, and it is thin. The data.org Digitisation of Oral Data playbook, produced with DSFSI, the University of Lagos, and Data Science Nigeria, is the only step-by-step African language-data playbook in the audit. Lacuna Fund's grantee guidance shapes proposals rather than daily work. Everything else in the layer is a paper.

The third layer is governance, and it is growing fast: the African Union's Continental AI Strategy and Data Policy Framework, the Malabo Convention, national data protection laws (Nigeria's NDPA, itself translated into Hausa, Yoruba, and Igbo; Kenya's DPA; South Africa's POPIA), and two community licensing frameworks: NOODL, from the Data Science Law Lab with DSFSI and CIPIT, and Esethu, from Lelapa AI and DSFSI. These instruments define what compliant, community-fair dataset work must look like, and no resource yet translates them into a workflow a project can follow.

The African wave has one shared weakness, and it applies to our own work first: 22 of the 34 African-specific resources are methodology papers bound to a single completed project, 27 are one-off publications or dormant, and none was written to be followed step by step. The knowledge is here. The playbook is not.

5 Reading the Gap Stage by Stage

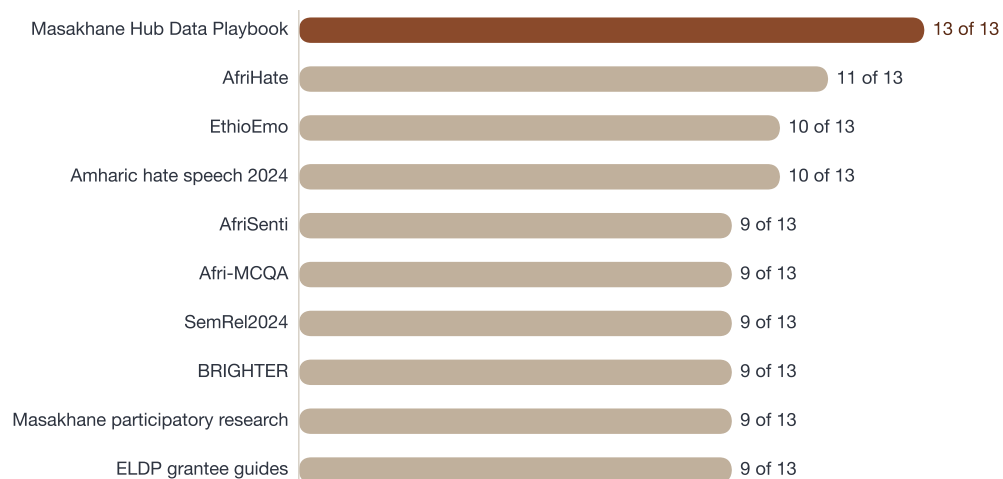
Crossing the four traditions against the 13 stages shows exactly what a new team can and cannot inherit today. Table 2 gives the coverage counts for the broadest resources, Figure 4 the same picture graphically, and Table 3 the per-stage breakdown for the proposed playbook against seven of the most instructive comparators.

Table 2: Lifecycle-stage coverage counts across the 13 criteria. Partial coverage is listed separately and not counted. Resources with report-author involvement are listed in Appendix B and scored under the same rules as everything else.

Resource	Stages fully covered (of 13)	Additional “Partial” entries
Masakhane Hub Data Playbook	13 (Planned)	0
AfriHate annotation methodology	11	2
EthioEmo annotation methodology	10	2
Amharic hate speech 2024 methodology	10	2
AfriSenti annotation methodology	9	3
Afri-MCQA multimodal QA methodology	9	4
SemRel2024 annotation methodology	9	3
BRIGHTER emotion datasets methodology	9	4
Masakhane participatory research framework	9	3
ELDP training resources and grantee guides	9	4
All other reviewed resources	0–8 each	varies

■ The broadest resources are research papers, and the ceiling is 11 of 13

Stages fully covered on the 13 comparison criteria; the proposed playbook row shows planned coverage. The median resource covers 4 of 13; resources with report-author involvement are listed in Appendix A.



Source: coverage matrices, Appendix A. Masakhane Hub Data Playbook entries are planned, not published.

Figure 4: Stages fully covered, for the proposed playbook and the nine broadest existing resources, counted from the Appendix A matrices.

Table 3: The 13 stages for the proposed playbook and seven comparators, simplified from the Appendix A matrices. ✓ fully covered; P = partial; ✗ not covered. Plan = planning; Comm = community engagement; Recr = recruitment; Pay = payment; Cons = consent and ethics; Lic = licensing; Dat = data collection (any modality); Guid = annotation guidelines; QC = quality control; Tool = tooling; Lang = scripts and language specifics; Ops = operations; Doc = documentation and publication. The playbook's ✓ marks are planned coverage, not published guidance.

Resource	Plan	Comm	Recr	Pay	Cons	Lic	Dat	Guid	QC	Tool	Lang	Ops	Doc
Masakhane Hub Data Playbook	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AfriHate	✓	✓	✓	✓	P	P	✓	✓	✓	✓	✓	✓	✓
EthioEmo	✓	P	✓	✓	✗	✓	✓	✓	✓	✓	✓	P	✓
Afri-MCQA	✓	P	✓	P	✓	✓	✓	✓	✓	P	P	✓	✓
BRIGHTER	✓	P	✓	✓	P	P	✓	✓	✓	✓	P	✓	✓
Masakhane participatory research	✓	✓	✓	✗	P	✓	✓	P	✓	✓	P	✓	✓
ELDP grantee guides	✓	P	P	P	✓	✓	✓	✓	P	✓	✓	✓	✓
data.org oral-data playbook	✓	✓	✓	P	✓	P	✓	P	✓	P	✓	P	P

Reading down the stages, four findings organize everything else:

- **The people stages remain the emptiest.** Recruitment is fully covered by 24 of 86 resources, payment by 8, and offline and field operations by 16. These are the stages practitioners rank hardest, and the covering resources are overwhelmingly project papers, ours included, rather than guidance a new team can follow. Nothing in the landscape explains cross-border payment mechanics or mobile money rails as a procedure.
- **The production stages live in different traditions that do not meet.** Data collection is fully covered by 48 of 86 resources, but Matrix 3 shows how uneven that number is across data types: speech recording appears in 19, image collection in 2, and OCR in 1. Annotation guidelines (27 of 86) and quality control (26 of 86) belong to the NLP methodology papers, while recording practice belongs to the fieldwork tradition; only 8 resources hold both sides. A team building an African speech dataset must today read two literatures and bridge them alone.
- **The release stages are well covered in general, but not translated into African practice.** Documentation and publication (45 of 86) and licensing (29 of 86) have mature frameworks behind them, and the African governance layer, from the AU instruments to NOODL and Esethu, now defines community-fair terms. What is missing is the connective guidance: which license for which project, what NDPA or POPIA compliance means for a speech corpus, how consent recorded in the field becomes metadata on a dataset hub.
- **Nothing is a playbook.** Of 86 resources, 11 are step-by-step playbooks, one addresses African language data, and 38 cover 3 stages or fewer. The ceiling of the whole landscape is 11 of 13, set by AfriHate, a research paper written to report a dataset, not to guide the next team, and it is an outlier: the median resource covers 4 stages. In practice a data creator must read several research papers and manuals side by side to assemble what one document should provide. Even the broadest pairing on paper, the AfriHate methodology

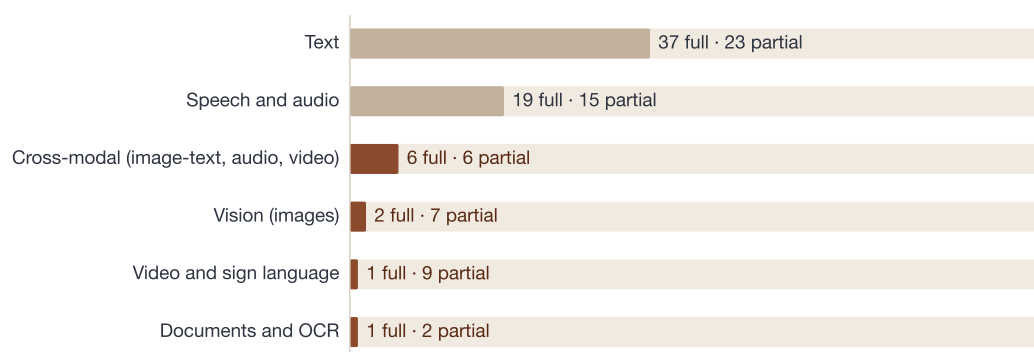
plus the ELDP fieldwork guides, spans the 13 stages only as a union: one is a hate speech research report, the other is archive-oriented field guidance, and nothing connects them into a single workflow.

5.1 Guidance stops at text and speech

African NLP is no longer only text and speech. Teams are building image datasets, digitising print and handwriting with OCR, recording sign language and gesture, and pairing images or audio with text. So we also checked which data types each of the 86 resources gives real guidance for, and the answer is short: text (37 of 86) and speech (19 of 86), then a cliff. Every full cross-modal entry in the audit (6 of 86) is a project paper from the African vision-language benchmark line: the Hausa Visual Genome, HaVQA, Afri-MCQA, CVQA, ArtELingo, and ArtELingo-28, several of them involving the report authors. Only two resources fully cover image dataset building (Afri-MCQA and CVQA, both documenting how annotators collected their own photographs), one covers document digitisation and OCR (SADiLaR's guidelines), and one covers sign-language video corpora (the Open Handbook's sign-language chapters, written for linguistics rather than machine learning).

Existing guidance stops at text and speech

Of 86 reviewed resources, how many give real guidance for building datasets of each data type



Source: playbook landscape audit of 86 resources, July 2026. All six full cross-modal entries are project papers, four of them by the report authors' teams; the proposed playbook plans tracks for all six groups.

Figure 5: Modality coverage across the 86 reviewed resources. Full detail is in Matrix 3, Appendix A.

This matters because the playbook we propose is planned around modality tracks, not text alone: text, speech and audio (including recognition, synthesis, and emotion), vision, documents and OCR (including handwritten text and page layout), video and sign language, and cross-modal tasks such as image captioning and visual question answering. The vision and cross-modal tracks will draw on the Hausa Visual Genome, HaVQA, and Afri-MCQA methodologies; for OCR, video, and sign language, the playbook's chapters will be the first guidance of their kind for African languages anywhere in this landscape.

6 The Playbook We Will Write

The landscape leaves a specific job undone. The practice exists, much of it in African project papers, ours among them, and no resource turns that practice into a path. The Masakhane

Hub Data Playbook will be a step-by-step, openly licensed guide to building African-language datasets, covering all 13 stages in one continuously maintained document and serving as the process companion to AfriAnnotate, the annotation platform proposed under the same grant.

Table 4 maps the playbook’s planned chapters to the stages they will close and the strongest existing material each will synthesize. The playbook will cite and build on that material rather than replace it; where a specialist source is better, the playbook will say so and link out.

Table 4: Planned chapter map of the Masakhane Hub Data Playbook. Stage abbreviations follow Table 3.

Planned chapter	Stages closed	Will synthesize, among others
Planning an African-language dataset project	Plan	Lacuna Fund grantee expectations; project structures from AfriSenti and BRIGHTER; pilot-first budgeting practice.
Partnering with the speaker community	Comm	Masakhane’s participatory framework; NaijaVoices’ data farming; the CARE principles; the Esethu benefit-sharing model.
Finding, training, and paying contributors	Recr, Pay	Recruitment and above-local-wage payment practice from AfriHate, EthioEmo, SemRel2024, and BRIGHTER; PAI’s fair-compensation baseline; mobile money and cross-border mechanics written as procedure for the first time.
Consent, licensing, and community ownership	Cons, Lic	NDPA, Kenya DPA, and POPIA compliance in practice; the AU Data Policy Framework; NOODL and Esethu alongside standard open licenses; consent templates usable with non-literate contributors.
Collecting speech, text, and other data under African conditions	Dat	The fieldwork tradition (ELDP, SIL BOLD); the data.org oral-data pipeline; Common Voice campaign practice; text-sourcing methods from NaijaSenti and AfriSenti and image-collection practice from Afri-MCQA and CVQA.
Writing guidelines and managing quality	Guid, QC	Annotated guideline excerpts from African projects (AfriHate’s three-level scheme, EthioEmo’s multi-label protocol, SemRel2024’s comparative annotation); agreement practice with honest thresholds.
Choosing tools and working offline	Tool, Ops	The companion 59-tool audit’s task-profile guidance; offline workflow discipline for phone-first teams; distributed-team practice from the African benchmark projects.
Documenting and publishing	Lang, Doc	Datasheets, Data Statements, and Croissant applied to African datasets; script and orthography documentation; archive deposit and Hub publication with consent and credit attached.
Modality tracks: vision, documents and OCR, video and sign language, cross-modal	All stages, per modality	The Hausa Visual Genome, HaVQA, and Afri-MCQA methodologies; the fieldwork tradition’s audio and video recording practice; the Open Handbook’s sign-language corpus chapters; SADiLaR’s digitisation standards; for African OCR and sign-language datasets, the first guidance of its kind.

Three commitments govern how the playbook will be written and kept.

It will synthesize, and it will credit. Every chapter will be grounded in named, cited practice: ours where ours is strongest, the fieldwork tradition’s where it is deeper, the governance instruments’ where they bind. The playbook will be the document that connects what the landscape already knows, not another resource standing on its own.

It will pair with AfriAnnotate without requiring it. The quality workflow, consent templates, licensing path, and publication pipeline will map directly onto the capabilities AfriAnnotate will implement, so a team will be able to follow the playbook with the tool for the smoothest path, or without it and lose nothing but convenience.

It will live. The playbook will be maintained as a versioned, openly governed document with graded contribution paths, refreshed against this report’s audit at each major revision, and translated into African languages, beginning with Hausa, Amharic, and Swahili. Twenty-seven of the 34 African-specific resources in our audit are one-off publications or dormant; the playbook’s sustainability model is designed against exactly that failure mode.

7 Conclusion

For well-resourced languages, dataset process knowledge lives inside long-standing institutions, and our audit contains them: the LDC has maintained corpus creation practice since 1992, ELRA published validation manuals, and the annotation methodology books teach their accumulated workflow. African-language teams have no equivalent institution to inherit from, and this report has measured what that absence looks like on paper: 86 resources a team might turn to, a coverage ceiling of 11 of 13 stages held by a single research paper, a median of 4, and the thinnest guidance exactly where practitioners report the most difficulty.

The measurement also shows where the strongest practice lives: in recent African annotation and benchmark projects, several of them led or co-led by the teams writing this report. That experience is our qualification, not our conclusion; our papers were scored in Appendix A under the same rules as everyone else’s. What no one has done, ourselves included, is turn that practice into a guide the next team can follow. The papers were written to report datasets; the playbook will be written to build them.

The Masakhane Hub Data Playbook is that missing institution, written down. It will gather the strongest existing practice, from ELDP’s recording standards to Masakhane’s participatory model to NOODL’s and Esethu’s community licensing to our own payment and quality workflows, write the chapters no one has written, and keep the whole current as a living, versioned, translated document paired with the AfriAnnotate platform.

This report is a deliverable of the Annotation and Playbook Grant, and it contributes the evidence base and a repeatable audit that can be rerun as the landscape changes, so the playbook’s coverage claims stay checkable as the work proceeds. The case it documents is simple: every African-language dataset project currently pays the cost of rediscovering its process, and that cost is paid again by the next project. The playbook will pay it once, and every team after that starts on chapter one instead of a blank page.

Appendix A: Full Coverage Matrices

Legend: ✓ = fully covered; P = partial; ✗ = not covered. Stage and modality abbreviations are defined in each caption; each resource’s African status and maintenance appear once, in the register (Table 5). The Masakhane Hub Data Playbook row records planned coverage, not published guidance; every other row is scored on published documents. Scoring rules and evidence notes are in Appendix B.

Table 5: Resource register: the 86 reviewed playbooks, handbooks, frameworks, methodology guides, and governance instruments. Resource names link to the canonical source.

Resource	Publisher / authors	Year	Type	African	Maintained
data.org Digitisation of Oral Data for NLP playbook	data.org with DSFSI (University of Pretoria), UNILAG, and Data Science Nigeria	2026	step-by-step playbook	Yes	one-off
data.org Africa CAN playbook	data.org and Africa CAN partners	2025-2026	step-by-step playbook	Yes	one-off
MasakhaNER annotation methodology	Adelani et al. (Masakhane)	2021-2022	guideline paper	Yes	dormant
MasakhaNEWS annotation methodology	Adelani et al. (Masakhane)	2023	guideline paper	Yes	one-off
Esethu framework and license	Rajab, Aremu, Marivate et al. (Lelapa AI / DSFSI, University of Pretoria)	2025	governance framework	Yes	active
DSFSI low-resource dataset creation approach (Setswana/Sepedi)	Marivate, Sefara et al. (DSFSI/CSIR)	2020	guideline paper	Yes	one-off
AI4D African Language Program methodology	Siminyu et al. (AI4D Africa)	2020-2021	guideline paper	Yes	one-off
Mozilla East Africa community datasets report	Mozilla Foundation with GIZ FAIR Forward	2023	guideline paper	Yes	one-off
SADiLaR digitisation guidelines	SADiLaR (North-West University / DSI)	2018-present	guideline paper	Yes	active
NaijaVoices data-farming methodology	Emezue et al. (NaijaVoices Community)	2024-2025	guideline paper	Yes	one-off
Kencorpus methodology	Wanjawa et al. (Kenyan universities)	2022-2023	guideline paper	Yes	dormant
Makerere East African benchmark lessons	Nakatumba-Nabende et al. (Makerere AI Lab)	2022	guideline paper	Yes	one-off
African Data Ethics framework	Abebe et al. (FAcCT)	2025	guideline paper	Yes	one-off
Deep Learning Indaba call for African datasets	Deep Learning Indaba	2023-2025	submission template	Yes	active
African Languages Lab collaborative approach	Issaka et al. (African Languages Lab)	2025	guideline paper	Yes	one-off
AU Continental AI Strategy	African Union	2024	policy strategy	Yes	active
AU Data Policy Framework	African Union	2022	policy framework	Yes	active
Nwulite Obodo Open Data License (NOODL)	Data Science Law Lab (University of Pretoria, Okorie) with DSFSI and CIPIT	2024	license framework	Yes	active
National data protection frameworks (NDPA, Kenya DPA, POPIA)	Nigeria DPC; ODPC Kenya; Information Regulator South Africa	2013-2023	legal framework	Yes	active
AU Malabo Convention	African Union	2014	legal framework	Yes	one-off
NaijaSenti annotation methodology	Muhammad et al. (HausaNLP / Masakhane)	2022	guideline paper	Yes	dormant
AfriSenti annotation methodology	Muhammad et al. (HausaNLP / EthioNLP / Masakhane)	2023	guideline paper	Yes	dormant
AfriHate annotation methodology	Muhammad et al. (HausaNLP / EthioNLP / Masakhane)	2025	guideline paper	Yes	one-off
EthioEmo annotation methodology	Belay, Azime, Ayele et al. (EthioNLP)	2025	guideline paper	Yes	one-off

Continued on next page

Resource register – continued

Resource	Publisher / authors	Year	Type	African	Maintained
EthioLLM / EthioBenchmark release methodology	Tonja, Azime, Belay et al. (EthioNLP)	2024	guideline paper	Yes	one-off
Amharic hate speech boundaries and intensities methodology	Ayele, Yimam et al. (EthioNLP / Bahir Dar / Univ. Hamburg)	2024	guideline paper	Yes	dormant
Amharic hate speech crowdsourcing line (RANLP / Toloka)	Ayele, Yimam, Belay et al. (EthioNLP)	2022-2023	guideline paper	Yes	dormant
Hausa Visual Genome (HaVG) methodology	Abdulmumin et al. (HausaNLP with UFAL / Idiap / KIIT)	2022	guideline paper	Yes	one-off
HaVQA visual question answering methodology	Parida, Abdulmumin et al. (HausaNLP with Silo AI / UFAL)	2023	guideline paper	Yes	one-off
Afri-MCQA multimodal cultural QA methodology	Tonja et al. (EthioNLP / HausaNLP / MBZUAI; ACL 2026 Best Social Impact Paper)	2026	guideline paper	Yes	one-off
SIL Comparative African Word List	Snider and Roberts (SIL)	2004	template	Yes	one-off
African Language Materials Archive (ALMA)	WARA/WARC and partners	2001	archive program	Yes	dormant
African Language Documentation LDC SP10	Seyfeddinipur (ed.), LDC	2016	guideline paper	Yes	one-off
Masakhane participatory research framework	Nekoto et al. (Masakhane)	2020	methodology framework	Yes	one-off
Lacuna Fund guidance suite	Lacuna Fund (Meridian Institute)	2020-2024	funder guidance	Partial	active
CLEAR Global / TWB language data initiative	CLEAR Global / Translators without Borders	2018-present	manual	Partial	active
SemRel2024 annotation methodology	Ousidhoum et al. (HausaNLP / EthioNLP / Masakhane collaborators)	2024	guideline paper	Partial	dormant
BRIGHTER emotion datasets methodology	Muhammad et al. (HausaNLP-led, incl. EthioNLP)	2025	guideline paper	Partial	one-off
CVQA culturally-diverse VQA methodology	Romero et al. (MBZUAI-led; EthioNLP and Masakhane contributors)	2024	guideline paper	Partial	one-off
ArtELingo affective captions methodology	Mohamed et al. (KAUST)	2022	guideline paper	Partial	dormant
ArtELingo-28 multilingual captions methodology	Mohamed et al. (KAUST/Oxford with HausaNLP's Ibrahim Said Ahmad)	2024	guideline paper	Partial	one-off
SIL BOLD methodology	SIL International (Reiman)	2010	step-by-step playbook	Partial	dormant
SIL Rapid Word Collection	SIL International (Moe et al.)	2010	step-by-step playbook	Partial	active
SIL Language and Culture Documentation Manual	Moeller, Self, Boerger, Reiman (SIL)	2016	manual	Partial	one-off
Understanding Linguistic Fieldwork	Meakins, Green, Turpin	2018	book	Partial	one-off
Open Handbook of Linguistic Data Management	Berez-Kroeker et al. (MIT Press)	2022	book	Partial	one-off
Check Before You Tech (FPCC)	First Peoples' Cultural Council	2020/2024	checklist playbook	Partial	active
Mozilla Common Voice community playbook	Mozilla Foundation	2019-present	step-by-step playbook	Partial	active

Continued on next page

Resource register – continued

Resource	Publisher / authors	Year	Type	African	Maintained
Google low-resource speech corpora methodology	Butryna, Kjartansson et al. (Google Research)	2018/2020	guideline paper	Partial	one-off
SIL SayMore documentation guides	SIL International	2015	manual	No	dormant
ELDP training resources and grantee guides	ELDP (BBAW Berlin)	2023	step-by-step playbook	No	active
ELAR deposit guidelines	Endangered Languages Archive	2023	documentation framework	No	active
UNESCO Language Vitality and Endangerment framework	UNESCO expert group (Brenzinger et al.)	2003	assessment framework	No	dormant
AIATSIS Code of Ethics	AIATSIS	2020	guideline paper	No	active
Paper and Talk manual (AIATSIS)	Thieberger (ed.)	1995	manual	No	dormant
FATSIL community protocols guide	FATSIL	2004	guideline paper	No	one-off
FPCC language toolkits	First Peoples' Cultural Council	2022	template	No	active
Essentials of Language Documentation	Gippert, Himmelmann, Mosel (eds.)	2006	book	No	one-off
Seven Dimensions of Portability	Bird and Simons	2003	guideline paper	No	one-off
CoLang workshop materials repository	CoLang Institute	2008-2020	training materials	No	active
Datasheets for Datasets	Geburu et al. (Microsoft Research)	2018/2021	documentation framework	No	one-off
Data Statements for NLP	Bender, Friedman, McMillan-Major (UW)	2018-2024	documentation framework	No	active
Data Cards Playbook	Google PAIR	2022	step-by-step playbook	No	dormant
Dataset Nutrition Label	Data Nutrition Project	2018/2022	documentation framework	No	active
Croissant metadata (MLCommons)	MLCommons working group	2024	documentation framework	No	active
FAIR Principles	Wilkinson et al. / GO FAIR	2016	principles framework	No	active
CARE Principles (GIDA)	Carroll et al. / Global Indigenous Data Alliance	2019/2020	principles framework	No	active
Te Mana Raraunga (Maori data sovereignty)	Te Mana Raraunga network	2016/2018	principles framework	No	active
Indigenous Protocol and AI Position Paper	Indigenous Protocol and AI Working Group	2020	guideline paper	No	one-off
Local Contexts TK/BC Labels	Local Contexts (Anderson, Christen)	2010s/2021	metadata labeling system	No	active
PAI Responsible Sourcing of Data Enrichment	Partnership on AI	2021/2022	guideline + templates	No	dormant
NeurIPS Datasets and Benchmarks checklist	NeurIPS	2021-present	release checklist	No	active
ACL Responsible NLP Research checklist	ACL Rolling Review	2022-present	release checklist	No	active
Hugging Face dataset documentation and cards	Hugging Face	2021-present	documentation framework	No	active
ODI Data Ethics Canvas	Open Data Institute	2019/2021	worksheet template	No	active
The Turing Way	Turing Way Community / Alan Turing Institute	2019-present	community handbook	No	active
IFRC Data Playbook	IFRC	2022	step-by-step playbook	No	active
Natural Language Annotation for ML (MATTER)	Pustejovsky and Stubbs (O'Reilly)	2012	book playbook	No	one-off

Continued on next page

Resource register – continued

Resource	Publisher / authors	Year	Type	African	Maintained
Handbook of Linguistic Annotation	Ide and Pustejovsky (Springer)	2017	book	No	one-off
Universal Dependencies contribution guidelines	UD consortium	2014-present	contribution playbook	No	active
ELRA validation guidelines	ELRA/ELDA	2000s	validation manual	No	dormant
LDC corpus creation practices	Linguistic Data Consortium	1992-present	institutional practices	No	active
Crowdsourcing best-practice guidelines (Sabou et al.)	Sabou, Bontcheva, Derczynski, Scharl (LREC)	2014	guideline paper	No	one-off
Google People + AI Guidebook (data chapters)	Google PAIR	2019/2024	design guidebook	No	active
Responsible Data Handbook	The Engine Room	2014/2016	manual	No	dormant
Snorkel programmatic labeling tutorials	Snorkel team	2019-2021	product tutorials	No	dormant

Table 6: Matrix 1. Project and people stages across the proposed Masakhane Hub Data Playbook and the 86 reviewed resources. Plan = planning; Comm = community engagement; Recr = recruitment; Pay = payment; Cons = consent and ethics; Lic = licensing and release; Ops = offline and field operations. African status and maintenance for every resource are recorded once, in the resource register (Table 5).

Resource	Plan	Comm	Recr	Pay	Cons	Lic	Ops
Masakhane Hub Data Playbook (proposed)	✓	✓	✓	✓	✓	✓	✓
data.org Digitisation of Oral Data for NLP playbook	✓	✓	✓	P	✓	P	P
data.org Africa CAN playbook	✓	P	P	✗	P	✗	P
MasakhaNER annotation methodology	P	P	✓	✗	P	P	P
MasakhaNEWS annotation methodology	P	P	✓	✗	P	P	P
Esethu framework and license	P	✓	P	✓	✓	✓	✗
DSFSI low-resource dataset creation approach (Setswana/Sepedi)	P	✗	✗	✗	✗	P	✗
AI4D African Language Program methodology	✓	P	✓	P	P	P	P
Mozilla East Africa community datasets report	✓	✓	✓	P	P	P	✓
SADiLaR digitisation guidelines	✗	✗	✗	✗	✗	P	✗
NaijaVoices data-farming methodology	P	✓	✓	P	✓	P	✓
Kencorpus methodology	P	P	P	✗	P	P	P
Makerere East African benchmark lessons	P	P	P	✗	P	P	P
African Data Ethics framework	✗	✓	✗	P	✓	P	✗
Deep Learning Indaba call for African datasets	P	P	✗	✗	P	P	✗
African Languages Lab collaborative approach	P	✓	P	✗	P	P	P
AU Continental AI Strategy	P	P	✗	✗	P	✗	✗
AU Data Policy Framework	P	P	✗	✗	P	P	✗
Nwulite Obodo Open Data License (NOODL)	P	P	✗	P	P	✓	✗
National data protection frameworks (NDPA, Kenya DPA, POPIA)	✗	✗	✗	✗	✓	P	P
AU Malabo Convention	✗	✗	✗	✗	P	✗	✗
NaijaSenti annotation methodology	✓	P	✓	✗	✗	P	P
AfriSenti annotation methodology	✓	P	✓	✗	P	P	✓
AfriHate annotation methodology	✓	✓	✓	✓	P	P	✓
EthioEmo annotation methodology	✓	P	✓	✓	✗	✓	P
EthioLLM / EthioBenchmark release methodology	P	P	✗	✗	✗	P	P
Amharic hate speech boundaries and intensities methodology	✓	P	✓	✓	✗	✓	P
Amharic hate speech crowdsourcing line (RANLP / Toloka)	✓	P	✓	P	✗	P	P
Hausa Visual Genome (HaVG) methodology	P	✗	P	✗	✗	✓	P
HaVQA visual question answering methodology	P	✗	✓	✗	✗	✓	P
Afri-MCQA multimodal cultural QA methodology	✓	P	✓	P	✓	✓	✓
SIL Comparative African Word List	✗	✗	✗	✗	✗	✗	✗
African Language Materials Archive (ALMA)	P	P	P	✗	✗	P	P
African Language Documentation LDC SP10	✓	✓	P	✗	P	✗	P
Masakhane participatory research framework	✓	✓	✓	✗	P	✓	✓
Lacuna Fund guidance suite	✓	✓	P	P	✓	✓	P
CLEAR Global / TWB language data initiative	P	P	✗	✗	P	P	P
SemRel2024 annotation methodology	✓	P	✓	✓	P	✗	✓
BRIGHTER emotion datasets methodology	✓	P	✓	✓	P	P	✓
CVQA culturally-diverse VQA methodology	✓	P	✓	✗	P	✓	P
ArtELingo affective captions methodology	P	✗	P	✗	✗	P	✗
ArtELingo-28 multilingual captions methodology	P	P	✓	✓	✗	P	✓
SIL BOLD methodology	✓	✓	P	✗	P	✗	✓

Continued on next page

Matrix 1 – continued

Resource	Plan	Comm	Recr	Pay	Cons	Lic	Ops
SIL Rapid Word Collection	✓	✓	✓	P	✗	✗	✓
SIL Language and Culture Documentation Manual	✓	✓	P	P	✓	P	✓
Understanding Linguistic Fieldwork	✓	✓	✓	P	✓	P	✓
Open Handbook of Linguistic Data Management	✓	✓	✗	✗	✓	✓	P
Check Before You Tech (FPCC)	✓	✓	✗	✗	✓	P	P
Mozilla Common Voice community playbook	P	✓	✓	✗	P	✓	P
Google low-resource speech corpora methodology	✓	P	✓	✗	P	✓	P
SIL SayMore documentation guides	P	✗	P	✗	✓	P	✓
ELDP training resources and grantee guides	✓	P	P	P	✓	✓	✓
ELAR deposit guidelines	P	P	✗	✗	✓	✓	P
UNESCO Language Vitality and Endangerment framework	✓	P	✗	✗	✗	✗	✗
AIATSIS Code of Ethics	✓	✓	P	P	✓	✓	P
Paper and Talk manual (AIATSIS)	✓	✓	P	✗	P	P	P
FATSIL community protocols guide	✓	✓	P	P	✓	✓	P
FPCC language toolkits	✓	✓	P	P	P	P	✓
Essentials of Language Documentation	✓	✓	P	P	✓	P	P
Seven Dimensions of Portability	P	✗	✗	✗	P	✓	✗
CoLang workshop materials repository	✓	✓	P	P	✓	P	P
Datasheets for Datasets	P	✗	✗	P	✓	✓	✗
Data Statements for NLP	P	P	P	P	P	P	✗
Data Cards Playbook	✓	P	✗	✗	P	P	✗
Dataset Nutrition Label	✗	✗	✗	✗	P	P	✗
Croissant metadata (MLCommons)	✗	✗	P	✗	P	✓	✗
FAIR Principles	P	✗	✗	✗	✗	✓	✗
CARE Principles (GIDA)	P	✓	✗	P	✓	P	✗
Te Mana Raraunga (Maori data sovereignty)	P	✓	✗	✗	✓	P	✗
Indigenous Protocol and AI Position Paper	P	✓	✗	✗	✓	P	✗
Local Contexts TK/BC Labels	P	✓	✗	✗	✓	✓	✗
PAI Responsible Sourcing of Data Enrichment	P	✗	P	✓	P	✗	P
NeurIPS Datasets and Benchmarks checklist	✗	✗	✗	P	P	✓	✗
ACL Responsible NLP Research checklist	✗	✗	P	P	✓	✓	✗
Hugging Face dataset documentation and cards	✗	✗	P	✗	P	✓	✗
ODI Data Ethics Canvas	✓	P	✗	✗	✓	P	✗
The Turing Way	✓	✓	✗	✗	✓	✓	P
IFRC Data Playbook	✓	✓	✗	✗	✓	P	P
Natural Language Annotation for ML (MATTER)	✓	✗	P	✗	✗	P	✗
Handbook of Linguistic Annotation	✓	✗	P	✗	✗	P	✗
Universal Dependencies contribution guidelines	P	P	✗	✗	✗	✓	✗
ELRA validation guidelines	P	✗	P	✗	P	✓	✗
LDC corpus creation practices	P	✗	P	P	P	✓	✗
Crowdsourcing best-practice guidelines (Sabou et al.)	✓	✗	✓	P	P	✗	✗
Google People + AI Guidebook (data chapters)	P	✗	P	✗	✓	✗	✗
Responsible Data Handbook	✓	P	✗	✗	✓	P	P
Snorkel programmatic labeling tutorials	P	✗	✗	✗	✗	✗	✗

Table 7: Matrix 2. Data, annotation, and publication stages across the proposed Masakhane Hub Data Playbook and the 86 reviewed resources. Dat = data collection in any modality (Matrix 3 gives the per-data-type breakdown of this column); Guid = annotation guidelines; QC = quality control; Tool = tooling guidance; Lang = scripts and language specifics; Doc = documentation and publication. African status and maintenance for every resource are recorded once, in the resource register (Table 5).

Resource	Dat	Guid	QC	Tool	Lang	Doc
Masakhane Hub Data Playbook (proposed)	✓	✓	✓	✓	✓	✓
data.org Digitisation of Oral Data for NLP playbook	✓	P	✓	P	✓	P
data.org Africa CAN playbook	✗	✗	✗	P	✗	P
MasakhaNER annotation methodology	✓	✓	✓	✓	P	P
MasakhaNEWS annotation methodology	✓	✓	✓	P	P	P
Esethu framework and license	P	✗	P	✗	P	✓
DSFSI low-resource dataset creation approach (Setswana/Sepedi)	✓	P	P	✗	P	P
AI4D African Language Program methodology	✓	✓	P	P	✓	✓
Mozilla East Africa community datasets report	✓	✗	P	P	✓	P
SADiLaR digitisation guidelines	✓	✗	P	P	P	P
NaijaVoices data-farming methodology	✓	P	✓	P	✓	P
Kencorpus methodology	✓	✓	P	P	✓	P
Makerere East African benchmark lessons	✓	✓	P	P	P	P
African Data Ethics framework	✗	✗	✗	✗	P	P
Deep Learning Indaba call for African datasets	P	✗	✗	✗	P	✓
African Languages Lab collaborative approach	✓	P	✓	P	✓	P
AU Continental AI Strategy	✗	✗	✗	✗	P	✗
AU Data Policy Framework	✗	✗	✗	✗	✗	✗
Nwulite Obodo Open Data License (NOODL)	✗	✗	✗	✗	P	P
National data protection frameworks (NDPA, Kenya DPA, POPIA)	P	✗	✗	✗	✗	P
AU Malabo Convention	✗	✗	✗	✗	✗	✗
NaijaSenti annotation methodology	✓	✓	✓	✓	✓	✓
AfriSenti annotation methodology	✓	✓	✓	✓	✓	✓
AfriHate annotation methodology	✓	✓	✓	✓	✓	✓
EthioEmo annotation methodology	✓	✓	✓	✓	✓	✓
EthioLLM / EthioBenchmark release methodology	✓	✗	✗	✗	P	✓
Amharic hate speech boundaries and intensities methodology	✓	✓	✓	✓	✓	✓
Amharic hate speech crowdsourcing line (RANLP / Toloka)	✓	P	✓	✓	✓	✓
Hausa Visual Genome (HaVG) methodology	✓	P	✓	✓	P	✓
HaVQA visual question answering methodology	✓	P	✓	✓	✓	✓
Afri-MCQA multimodal cultural QA methodology	✓	✓	✓	P	P	✓
SIL Comparative African Word List	✓	P	P	✗	✓	P
African Language Materials Archive (ALMA)	✓	✗	✗	P	✓	✓
African Language Documentation LDC SP10	✓	P	P	P	✓	P
Masakhane participatory research framework	✓	P	✓	✓	P	✓
Lacuna Fund guidance suite	P	P	P	✗	P	✓
CLEAR Global / TWB language data initiative	P	✗	✗	✗	✓	P
SemRel2024 annotation methodology	✓	✓	✓	P	✓	✓
BRIGHTER emotion datasets methodology	✓	✓	✓	✓	P	✓
CVQA culturally-diverse VQA methodology	✓	✓	✓	P	P	✓
ArtELingo affective captions methodology	✓	P	P	P	P	✓
ArtELingo-28 multilingual captions methodology	✓	✓	✓	P	P	✓
SIL BOLD methodology	✓	P	P	✓	P	P
SIL Rapid Word Collection	✓	P	P	✓	P	P

Continued on next page

Matrix 2 – continued

Resource	Dat	Guid	QC	Tool	Lang	Doc
SIL Language and Culture Documentation Manual	✓	✓	P	✓	P	✓
Understanding Linguistic Fieldwork	✓	✓	P	P	P	P
Open Handbook of Linguistic Data Management	✓	P	✓	✓	P	✓
Check Before You Tech (FPCC)	P	✗	✗	P	✓	✗
Mozilla Common Voice community playbook	✓	P	P	P	✓	P
Google low-resource speech corpora methodology	✓	P	✓	P	✓	P
SIL SayMore documentation guides	✓	P	P	✓	✗	✓
ELDP training resources and grantee guides	✓	✓	P	✓	✓	✓
ELAR deposit guidelines	P	✗	P	✓	✗	✓
UNESCO Language Vitality and Endangerment framework	✗	✗	P	✗	P	P
AIATSI Code of Ethics	✗	✗	P	✗	✗	✓
Paper and Talk manual (AIATSI)	✓	✓	P	P	✓	✓
FATSIL community protocols guide	P	✗	✗	✗	✗	P
FPCC language toolkits	P	✗	✗	✓	✓	P
Essentials of Language Documentation	✓	✓	P	P	✓	✓
Seven Dimensions of Portability	P	P	P	✓	P	✓
CoLang workshop materials repository	✓	✓	✗	✓	P	✓
Datasheets for Datasets	✗	P	P	✗	✗	✓
Data Statements for NLP	P	P	P	✗	✓	✓
Data Cards Playbook	✗	P	P	✗	✗	✓
Dataset Nutrition Label	✗	✗	P	✗	✗	✓
Croissant metadata (MLCommons)	✗	P	P	✗	P	✓
FAIR Principles	✗	✗	✗	✗	✗	✓
CARE Principles (GIDA)	✗	✗	✗	✗	✗	P
Te Mana Rauunga (Maori data sovereignty)	✗	✗	P	✗	P	P
Indigenous Protocol and AI Position Paper	✗	✗	✗	✗	P	✗
Local Contexts TK/BC Labels	✗	✗	✗	✗	P	✓
PAI Responsible Sourcing of Data Enrichment	✗	✓	P	P	✗	✗
NeurIPS Datasets and Benchmarks checklist	✗	P	P	✗	✗	✓
ACL Responsible NLP Research checklist	✗	P	P	✗	P	✓
Hugging Face dataset documentation and cards	✓	P	✗	✗	P	✓
ODI Data Ethics Canvas	✗	✗	P	✗	✗	P
The Turing Way	✗	✗	P	P	✗	✓
IFRC Data Playbook	P	✗	P	P	✗	P
Natural Language Annotation for ML (MATTER)	✓	✓	✓	P	✗	P
Handbook of Linguistic Annotation	✓	✓	✓	✓	P	P
Universal Dependencies contribution guidelines	P	✓	✓	P	✓	✓
ELRA validation guidelines	✓	P	✓	✗	P	✓
LDC corpus creation practices	✓	P	P	P	P	✓
Crowdsourcing best-practice guidelines (Sabou et al.)	P	✓	✓	P	✗	✗
Google People + AI Guidebook (data chapters)	P	✓	P	P	✗	P
Responsible Data Handbook	P	✗	✗	P	✗	P
Snorkel programmatic labeling tutorials	✓	P	P	✓	✗	✗

Table 8: Matrix 3. The per-data-type breakdown of data collection: which data types each resource gives real guidance for. A resource's Dat cell in Matrix 2 is covered exactly when at least one group here is covered. Txt = text; Sp = speech and audio; Vis = vision (images); OCR = documents and OCR; VSL = video and sign language; XM = cross-modal (image-text, aligned audio, video, and text). African status and maintenance for every resource are recorded once, in the resource register (Table 5).

Resource	Txt	Sp	Vis	OCR	VSL	XM
Masakhane Hub Data Playbook (proposed)	✓	✓	✓	✓	✓	✓
data.org Digitisation of Oral Data for NLP playbook	P	✓	✗	✗	✗	P
data.org Africa CAN playbook	✗	✗	✗	✗	✗	✗
MasakhaNER annotation methodology	✓	✗	✗	✗	✗	✗
MasakhaNEWS annotation methodology	✓	✗	✗	✗	✗	✗
Esethu framework and license	P	P	✗	✗	✗	✗
DSFSI low-resource dataset creation approach (Setswana / Sepedi)	✓	✗	✗	✗	✗	✗
AI4D African Language Program methodology	✓	P	✗	✗	✗	✗
Mozilla East Africa community datasets report	P	✓	✗	✗	✗	✗
SADiLaR digitisation guidelines	✓	P	✗	✓	✗	✗
NaijaVoices data-farming methodology	✓	✓	✗	✗	✗	✗
Kencorpus methodology	✓	P	✗	✗	✗	✗
Makerere East African benchmark lessons	✓	✓	✗	✗	✗	✗
African Data Ethics framework	✗	✗	✗	✗	✗	✗
Deep Learning Indaba call for African datasets	P	P	✗	✗	✗	✗
African Languages Lab collaborative approach	✓	✓	✗	✗	✗	✗
AU Continental AI Strategy	✗	✗	✗	✗	✗	✗
AU Data Policy Framework	✗	✗	✗	✗	✗	✗
Nwulite Obodo Open Data License (NOODL)	✗	✗	✗	✗	✗	✗
National data protection frameworks (NDPA, Kenya DPA, POPIA)	P	P	✗	✗	✗	✗
AU Malabo Convention	✗	✗	✗	✗	✗	✗
NaijaSenti annotation methodology	✓	✗	✗	✗	✗	✗
AfriSenti annotation methodology	✓	✗	✗	✗	✗	✗
AfriHate annotation methodology	✓	✗	✗	✗	✗	✗
EthioEmo annotation methodology	✓	✗	✗	✗	✗	✗
EthioLLM / EthioBenchmark release methodology	✓	✗	✗	✗	✗	✗
Amharic hate speech boundaries and intensities methodology	✓	✗	✗	✗	✗	✗
Amharic hate speech crowdsourcing line (RANLP / Toloka)	✓	✗	✗	✗	✗	✗
Hausa Visual Genome (HaVG) methodology	✓	✗	P	✗	✗	✓
HaVQA visual question answering methodology	✓	✗	P	✗	✗	✓
Afri-MCQA multimodal cultural QA methodology	✓	✓	✓	✗	✗	✓
SIL Comparative African Word List	✓	✗	✗	✗	✗	✗
African Language Materials Archive (ALMA)	✓	✗	✗	P	✗	✗
African Language Documentation LDC SP10	P	✓	✗	✗	✗	✗
Masakhane participatory research framework	✓	✗	✗	✗	✗	✗
Lacuna Fund guidance suite	P	P	P	✗	✗	✗
CLEAR Global / TWB language data initiative	P	✗	✗	✗	✗	✗
SemRel2024 annotation methodology	✓	✗	✗	✗	✗	✗
BRIGHTER emotion datasets methodology	✓	✗	✗	✗	✗	✗
CVQA culturally-diverse VQA methodology	✓	✗	✓	✗	✗	✓
ArtELingo affective captions methodology	✓	✗	P	✗	✗	✓
ArtELingo-28 multilingual captions methodology	✓	✗	P	✗	✗	✓
SIL BOLD methodology	✗	✓	✗	✗	✗	✗
SIL Rapid Word Collection	✓	✗	✗	✗	✗	✗

Continued on next page

Matrix 3 – continued

Resource	Txt	Sp	Vis	OCR	VSL	XM
SIL Language and Culture Documentation Manual	<i>P</i>	✓	✗	✗	<i>P</i>	✗
Understanding Linguistic Fieldwork	✓	✓	✗	✗	<i>P</i>	✗
Open Handbook of Linguistic Data Management	<i>P</i>	<i>P</i>	✗	✗	✓	✗
Check Before You Tech (FPCC)	✗	<i>P</i>	✗	✗	✗	✗
Mozilla Common Voice community playbook	<i>P</i>	✓	✗	✗	✗	✗
Google low-resource speech corpora methodology	<i>P</i>	✓	✗	✗	✗	✗
SIL SayMore documentation guides	✗	✓	✗	✗	<i>P</i>	✗
ELDP training resources and grantee guides	<i>P</i>	✓	✗	✗	<i>P</i>	<i>P</i>
ELAR deposit guidelines	<i>P</i>	<i>P</i>	✗	✗	✗	✗
UNESCO Language Vitality and Endangerment framework	✗	✗	✗	✗	✗	✗
AIATSI Code of Ethics	✗	✗	✗	✗	✗	✗
Paper and Talk manual (AIATSI)	✓	✗	✗	<i>P</i>	✗	✗
FATSIL community protocols guide	✗	<i>P</i>	✗	✗	✗	✗
FPCC language toolkits	<i>P</i>	<i>P</i>	✗	✗	<i>P</i>	✗
Essentials of Language Documentation	✓	✓	✗	✗	<i>P</i>	<i>P</i>
Seven Dimensions of Portability	<i>P</i>	<i>P</i>	✗	✗	✗	✗
CoLang workshop materials repository	<i>P</i>	✓	✗	✗	<i>P</i>	✗
Datasheets for Datasets	✗	✗	✗	✗	✗	✗
Data Statements for NLP	<i>P</i>	<i>P</i>	✗	✗	✗	✗
Data Cards Playbook	✗	✗	✗	✗	✗	✗
Dataset Nutrition Label	✗	✗	✗	✗	✗	✗
Croissant metadata (MLCommons)	✗	✗	✗	✗	✗	✗
FAIR Principles	✗	✗	✗	✗	✗	✗
CARE Principles (GIDA)	✗	✗	✗	✗	✗	✗
Te Mana Raraunga (Maori data sovereignty)	✗	✗	✗	✗	✗	✗
Indigenous Protocol and AI Position Paper	✗	✗	✗	✗	✗	✗
Local Contexts TK/BC Labels	✗	✗	✗	✗	✗	✗
PAI Responsible Sourcing of Data Enrichment	✗	✗	✗	✗	✗	✗
NeurIPS Datasets and Benchmarks checklist	✗	✗	✗	✗	✗	✗
ACL Responsible NLP Research checklist	✗	✗	✗	✗	✗	✗
Hugging Face dataset documentation and cards	✓	<i>P</i>	<i>P</i>	✗	✗	<i>P</i>
ODI Data Ethics Canvas	✗	✗	✗	✗	✗	✗
The Turing Way	✗	✗	✗	✗	✗	✗
IFRC Data Playbook	<i>P</i>	✗	✗	✗	✗	✗
Natural Language Annotation for ML (MATTER)	✓	✗	✗	✗	✗	✗
Handbook of Linguistic Annotation	✓	✓	✗	✗	<i>P</i>	<i>P</i>
Universal Dependencies contribution guidelines	<i>P</i>	✗	✗	✗	✗	✗
ELRA validation guidelines	✓	✓	✗	✗	<i>P</i>	<i>P</i>
LDC corpus creation practices	✓	✓	✗	✗	✗	✗
Crowdsourcing best-practice guidelines (Sabou et al.)	<i>P</i>	✗	✗	✗	✗	✗
Google People + AI Guidebook (data chapters)	<i>P</i>	✗	<i>P</i>	✗	✗	✗
Responsible Data Handbook	<i>P</i>	✗	✗	✗	✗	✗
Snorkel programmatic labeling tutorials	✓	✗	✗	✗	✗	✗

Appendix B: Scoring Rules and Evidence References

Scoring rules

Each resource was scored from its public documents at the report date. A cell is scored ✓ (the resource gives substantive, actionable guidance on the stage), *P* (the stage is touched but not

usably treated), or ✗ (the stage is absent). A ✓ records that usable guidance exists, not that it targets African languages; the African column records specificity separately. Resources are listed African-specific first, then partially African, then generic. The Masakhane Hub Data Playbook row records planned coverage, not published guidance; every other row, including the report authors' own publications, is scored on published documents under the same rules. The maintenance column records whether the guidance itself is maintained. A research paper is one-off by nature, even when its dataset or repository remains active; active is reserved for resources whose guidance is updated, such as living handbooks, operating licenses, and recurring calls.

Legend: ✓ = fully covered; P = partial; ✗ = not covered. African: Yes / Partial / No. Maintenance: active, dormant, or one-off publication, as at the report date. Stage abbreviations follow Table 3. Matrix 1 and Matrix 2 cover the 14 lifecycle stages; Matrix 3 covers the six data-type groups from Section 5.1.

Evidence references

- **Resource register:** every resource name in Table 5 links to its canonical public source; scoring was performed against those documents at the report date.
- **Companion study:** the practitioner survey (n = 26), the 96-paper dataset audit, and the 59-tool capability audit are documented in Gap Analysis of Annotation Tools for AfricaNLP (this grant, July 2026); Figures 1 and 2 are reproduced from it.
- **The report authors' own resources:** NaijaSenti (LREC 2022), AfriSenti (2023, with SemEval-2023 Task 12), AfriHate (NAACL 2025), SemRel2024 (Findings of ACL 2024, with SemEval-2024 Task 1), BRIGHTER (ACL 2025, Best Resource Paper, with SemEval-2025 Task 11), EthioEmo (COLING 2025), EthioLLM (LREC-COLING 2024), the Amharic hate speech corpora (RANLP 2023; 2024), the Hausa Visual Genome (LREC 2022), HaVQA (Findings of ACL 2023), and Afri-MCQA (ACL 2026, Best Social Impact Paper) were scored under the same rules as every other resource, from their published papers and repositories.
- **data.org playbooks:** the Digitisation of Oral Data for NLP of Low-Resource Languages playbook was produced by data.org with DSFSI (University of Pretoria), the University of Lagos, and Data Science Nigeria; data.org also publishes the Africa CAN capacity-building playbook, which we score separately.
- **Governance instruments:** the AU Continental AI Strategy and Data Policy Framework, the Malabo Convention, national data protection statutes (NDPA, Kenya DPA, POPIA), and the NOODL and Esethu licensing frameworks are scored as guidance a dataset team would consult, not summarized as law; teams should verify current legal requirements with counsel.
- **Scope caveat:** the audit covers resources locatable through public search in English and French at the report date; internal project handbooks that organizations have not published could not be scored, so landscape counts are lower bounds on what exists but accurate for what a new team can actually find and use.

Appendix C: Abbreviations

Term	Meaning
AI4D	Artificial Intelligence for Development (Africa), an IDRC/Sida-supported program that funded African language dataset challenges and fellowships.
AU	African Union; its Continental AI Strategy (2024) and Data Policy Framework (2022) are scored in the audit.
CARE	Collective Benefit, Authority to Control, Responsibility, and Ethics: principles for Indigenous data governance.
Croissant	A machine-readable metadata format for ML datasets, developed by an ML-Commons working group.
DSFSI	Data Science for Social Impact research group, University of Pretoria.
ELDP / ELAR	Endangered Languages Documentation Programme and the Endangered Languages Archive.
Esethu	A dataset governance framework and license with revenue reinvestment for data creators (Lelapa AI / DSFSI, 2025).
FAIR	Findable, Accessible, Interoperable, Reusable: principles for scientific data management.
HF Hub	The Hugging Face Hub, the primary public distribution platform for machine learning datasets and models.
IAA	Inter-annotator agreement.
IFRC	International Federation of Red Cross and Red Crescent Societies.
IRB	Institutional review board (research ethics committee).
NDPA / POPIA	Nigeria Data Protection Act 2023; South Africa's Protection of Personal Information Act.
NLP	Natural language processing.
NOODL	Nwulite Obodo Open Data License: a tiered license for African datasets (Data Science Law Lab, University of Pretoria, with DSFSI and CIPIT, 2024).
PAI	Partnership on AI.
QC	Quality control.
SADiLaR	South African Centre for Digital Language Resources.
SIL	SIL International, a language development organization with a long methodology tradition.
TWB	Translators without Borders (now CLEAR Global).
UNILAG	University of Lagos, Nigeria.