Masakhane Playbook

Democratizing machine translation for African languages

Table of contents:

- Key Components of Planning
  - Budgeting Annotation Costs
  - Time Estimation
  - Scaling Strategy

Data Cleaning and Preprocessing

- Why Data Cleaning Matters
- Key Steps in Data Cleaning
  - Deduplication, Normalization, and Filtering
  - Language Detection and Formatting
  - Noise and Toxicity Handling
  - Missing and Corrupted Data Handling

Data Provenance and Traceability

- Why Provenance Matters
- Key Components of Provenance
  - Source Tracking
  - Data Lineage
  - Transformation Logs

Ethics, Bias, and Governance

- Why Ethics and Governance Matter
- Key Components of Ethical Data Practices
  - Bias Identification and Mitigation
  - PII Detection and Removal
  - Anonymization Strategies
  - Sensitive Attribute and Content Handling
  - Fair Representation
  - Risk Documentation and Transparency
- 📄Annotation Task Design and Human Factors
- 📄Inclusive and Bias-Aware Annotation
- 📄Training and Guidelines
- 📄Workflow and Adjudication

Annotation Task Design and Human Factors

- Task complexity estimation
- Annotator fatigue
- UI/UX considerations

Inclusive and Bias-Aware Annotation

- Why Inclusion and Bias Awareness Matter
- Key Components
  - Gender, Age, and Cultural Diversity
  - Recording Annotator Personas

- Bias Awareness Training
- Community Participation

Training and Guidelines

- Annotation Guidelines with Examples
- Training and Calibration Rounds
- Pilot Annotation and Iteration
- Minimum Viable Dataset per Task

Workflow and Adjudication

- Multi-annotator setup
- Task assignment and redundancy
- Disagreement resolution and expert adjudication
- 📄 Data Quality Management
- 📄 Data Quality Assurance

Data Quality Management

- Class imbalance handling
- Outlier and noise detection
- Error analysis pipelines
- Dataset versioning and updates

Data Quality Assurance

- Inter-Annotator Agreement (Kappa, Alpha)
- Gold data and control checks
- Auditing and spot-checking
- Feedback loops
- 📄 Image Data
- 📄 Multimodal Data
- 📄 Speech Data
- 📄 Text Data

Image Data

Multimodal Data

Speech Data

Text Data

- 📄 Templates & Artifacts
- 📄 Consent Form Template
- 📄 Contributor Agreement Template
- 📄 Data Ownership Documentation Template
- 📄 Licensing and Compliance
- 📄 Sustainability Plan Template
- 📄 Documentation and Reporting
- 📄 Tooling and Infrastructure

1. Introduction

A comprehensive guide to dataset design, annotation, and task formulation for building reliable and responsible language AI systems.

> 📚 **CITING THIS PLAYBOOK**
>
> Using this resource in research, teaching, or a project? **Jump to the citation block** at the bottom of this page for BibTeX, APA, and other formats. The full citation page lives at `/cite`.

> 🤝 **HELP BUILD THE PLAYBOOK**
>
> This is a **community-driven** resource. If you spot a gap, want to write a chapter, translate a page, or suggest an improvement — contributions from researchers, practitioners, students, and language experts are very welcome. See the **contribution guide** to get started, or join the conversation on Discord.

Welcome to the dataset design and annotation playbook!

This playbook will help you plan and develop **training and evaluation datasets**, define **annotation schemas**, and design **AI tasks** across different languages, domains, and modalities. It provides guidance on dataset structuring, labeling strategies, and ethical considerations for language technologies.

How to read this playbook

The playbook is organised end-to-end through the dataset lifecycle, but you don't have to read it linearly. Pick the path that fits where you are:

- **New to dataset design.** Start here, then read **chapters 2–4** in order — Data Collection → Annotation Design → Data Quality. They build on each other and cover the foundations everyone needs.

- **You already have raw data, want guidance on annotation.** Jump to **chapter 3 (Annotation Design and Workforce Management)**, then **chapter 4 (Data Quality Assurance and Validation)**.
- **You're working with a specific modality** (speech, multimodal, low-resource scripts). Skip to **chapter 5 (Modality-Specific Task Design)**.
- **You're using LLMs to generate or augment data.** Read **chapter 7 (LLM-Assisted and Synthetic Data Generation)** for the trade-offs and safeguards.
- **You're preparing a dataset for release or publication.** Read **chapter 6 (Documentation, Data Release, and Governance)** and **chapter 9 (Dataset Lifecycle Management and Release Checklist)**.
- **You're a coordinator onboarding a team or community group.** See Onboarding a Team and Running a Playbook Workshop.
- **You're reading offline or on a slow connection.** Use **Download PDF** in the navbar — the entire playbook bundles into a single file, regenerated automatically on every release.
- **You'd rather read in Hausa, Amharic, Swahili, French, or Portuguese.** Use the language switcher in the top-right of the navbar. Translations are community-maintained and grow over time.

Throughout the playbook, you'll find practical templates (consent forms, annotation guidelines, governance checklists), worked examples from real African-language projects, and links to source datasets and tools you can reuse.

Who is this playbook for?

This playbook is designed for:

- **Researchers** working on NLP dataset creation and evaluation
- **Annotation teams** developing labeled datasets
- **Project managers and coordinators** overseeing data collection and annotation workflows
- **AI practitioners** designing and evaluating language models
- **Students and academics** studying dataset design and annotation
- **Multilingual communities** contributing to language resources

- **Trainers and facilitators** who run workshops or onboarding sessions for contributors

What will you learn?

By the end of this playbook, you will understand:

- How to define the **purpose and scope** of a dataset
- Differences between **training and evaluation datasets**
- Trade-offs between **scale and quality**
- How to design **label schemas and ontologies**
- Approaches for **multi-label, single-label, and structured outputs**
- How to handle **ambiguity, edge cases, and annotation boundaries**
- Best practices for **multilingual and cross-lingual dataset design**
- Ethical considerations, risks, and limitations in dataset creation

How to use this playbook

Each section of this playbook contains:

- **Clear explanations** of dataset design principles
- **Structured guidance** for task and schema definition
- **Examples and edge cases** to support annotation decisions
- **Practical recommendations** for dataset creation workflows
- **Ethical considerations** to guide responsible use

Getting Started

Ready to begin? Start with our foundational sections:

1. **Purpose of this Playbook** – Understand target users, scope, and intended use
2. **How to Use This Playbook** – Learn how to navigate chapters and contribute
3. **Dataset Types and Design Goals** – Explore dataset categories and trade-offs
4. **Task and Schema Definition** – Define tasks, labels, and annotation structures

5. **Glossary and Terminology** – Learn key concepts and definitions

Purpose of this playbook

- Target users and communities
- Languages, domains, and modalities covered
- Intended use and risks

Dataset Types and Design Goals

- Training vs evaluation datasets
- General-purpose vs domain-specific datasets
- Scale vs quality trade-offs
- Monolingual, multilingual, cross-lingual setups

Task and Schema Definition

- Task formulation (classification, generation, alignment, retrieval)
- Label schema and ontology design
- Multi-label vs single-label vs structured outputs
- Ambiguity, edge cases, and annotation boundaries

Glossary and Terminology

A reference section providing clear definitions of the key terms used throughout the playbook — see the **Glossary** for definitions of *annotation, inter-annotator agreement, Cohen's kappa, low-resource language, modality,* and other terms.

How to cite this playbook

If the Masakhane Playbook informs your research, teaching, or project, please cite it.

BibTeX:

```
@misc{masakhane2026playbook,
  author       = {{Masakhane Community}},
  title        = {Masakhane Playbook: A Practical Guide for Building NLP
Systems for African Languages},
  year         = {2026},
  publisher    = {Masakhane},
  url          = {https://masakhanehubnlp.github.io/MasakhanePlaybook/},
  note         = {Open-source community resource}
}
```

Plain text (APA-style):

> Masakhane Community. (2026). *Masakhane Playbook: A Practical Guide for Building NLP Systems for African Languages*. https://masakhanehubnlp.github.io/MasakhanePlaybook/

For other formats (MLA, Chicago, etc.) and a machine-readable `CITATION.cff`, see the /cite page.

If you reference a specific chapter, please include the chapter title and its URL.

Cite this page 4 min read ✦ Contribute

*Last updated on **May 2, 2026** by **Seid Muhie Yimam***

Onboarding a Team

This page is for coordinators and project leads who want to introduce the playbook to a new team or community group.

Where to Start by Role

Rather than sharing the full playbook with a new team, start with the chapter most relevant to their immediate task.

Role	Recommended starting point
Annotator	Training and Guidelines, then the relevant modality chapter
Voice recorder	Speech Data
Reviewer / quality checker	Data Quality Assurance and Validation
Coordinator	Cost and Resource Planning → Annotation Design
Linguist	Annotation Design → Inclusive and Bias-Aware Design
Dataset release lead	Documentation and Governance → Dataset Lifecycle

First-Session Structure

A single 60–90 minute orientation session is usually enough to get a team started:

1. Walk through the relevant chapter together (screen share or printed copy)
2. Work through one concrete example from the actual project — not an abstract sample
3. Run a short practice task and discuss any questions before independent work begins

Onboarding Materials to Prepare

- [ ] Link (or printout) of the relevant chapter
- [ ] 3–5 real annotation or recording examples from your project
- [ ] Contact details for who to reach with tool or task questions
- [ ] Offline copy of materials for contributors without reliable internet (see Inclusive Access)

Checking Readiness

Before a contributor starts independent work, confirm they can:

- Navigate to the relevant playbook section without help
- Correctly identify the label or action for at least three practice examples
- Reach their point of contact if something is unclear

Contributors who cannot pass a short readiness check benefit from a second orientation rather than starting at full scale.

*Last updated on **May 2, 2026** by **Seid Muhie Yimam***

Running a Playbook Workshop

For larger groups or project kick-offs, a structured workshop helps align a team on guidelines and workflows before annotation begins.

When a Workshop Is Worth It

A dedicated session makes sense when:

- You are starting a new project with contributors who have not used the playbook before
- You need to align a large team on common terminology before annotation begins
- You are piloting the playbook in a new language community and want to collect localization feedback

For small teams or individual onboarding, a one-to-one walkthrough is usually sufficient — see Onboarding a Team.

Workshop Formats

Format	Best for	Duration
Orientation session	Introduce the playbook to a new team	2–3 hours
Task deep dive	Train contributors on one chapter (e.g., speech recording, text annotation)	Half day
Full onboarding	Bring all project roles together before a project launch	Full day
Feedback and localization	Review a chapter with community experts, collect revision input	2–3 hours

Preparing a Session

1. **Identify the relevant chapters** — limit to 1–2 per session; do not attempt the full playbook at once
2. **Prepare a concrete worked example** — use real samples from your project, not abstract examples
3. **Distribute materials in advance** — share the chapter link or a printout at least 3 days beforehand
4. **Assign roles** — designate a facilitator, a note-taker, and a timekeeper before the session begins
5. **Make materials available offline** — see Inclusive Access

A Simple Session Structure

```
00:00 – 00:15  Welcome and objectives
00:15 – 00:30  Playbook overview (structure, how to navigate)
00:30 – 01:00  Deep dive into the relevant chapter (facilitator-led)
01:00 – 01:30  Hands-on task in small groups (3–5 people)
01:30 – 01:50  Group debrief (what was clear, what was confusing)
01:50 – 02:00  Next steps and Q&A
```

For a full-day workshop, repeat the deep-dive and hands-on blocks for each additional chapter with breaks between.

Facilitation Tips

- Start with a concrete task, not a lecture — participants engage faster when they have something to do
- **Capture disagreements** — where participants interpret guidelines differently is where the playbook needs more clarity; document these moments and submit them as feedback
- Time-box discussions; have someone keep the session on schedule

After the Workshop

- Share a written summary within 48 hours: what was covered, what questions arose, what next steps are
- Log any sections that caused confusion — these are candidates for improvement
- Submit feedback via the GitHub repository or the built-in feedback form on the site
- Identify one or two participants who can serve as local playbook champions — people who can answer questions and onboard future contributors independently

*Last updated on **May 2, 2026** by **Seid Muhie Yimam***

Inclusive Access

The playbook should be usable by people with different levels of technical background, literacy, language, and connectivity. This page guides coordinators on adapting it for diverse audiences.

Adapting for Non-Technical Audiences

Many contributors — annotators, voice recorders, translators — are not researchers. When introducing the playbook to them:

- Share only the sections relevant to their task; do not link to the full playbook
- Replace technical terms with plain-language equivalents, or walk through the Glossary first
- Demonstrate tasks live (screen share or in-person) before asking contributors to read independently
- Use concrete examples from the contributor's own language and domain

Adapting for Low-Literacy Contributors

- Convert key instructions into short numbered steps with screenshots or illustrations
- Offer oral walkthroughs as an alternative to written reading — a coordinator reads through the section and takes questions
- Prepare a printed quick-reference card (A5, one side) with the most important steps for the specific task
- Test materials with community members before finalizing — contributors with low literacy often surface genuine clarity problems that polished prose hides

Gender-Inclusive Facilitation

- Actively invite women and underrepresented participants to contribute examples and ask questions

- Use training examples that reflect diverse speakers, topics, and perspectives — avoid stereotyped scenarios
- Offer flexible session timing to accommodate caregiving responsibilities
- Track participation by gender and adjust facilitation if one group dominates

Multilingual Access

- Translate the sections relevant to your project before the training session — do not rely on machine translation for consent forms or contributor agreements
- Maintain a local glossary of key annotation terms in the community's working language
- Invite a bilingual co-facilitator for sessions where participants have limited English

Offline and Low-Connectivity Access

For contributors working without reliable internet:

- Use **Download PDF** in the navbar to get the full playbook as a single file — suitable for printing or sharing via WhatsApp or USB
- For workshops in low-connectivity venues, pre-download or print only the relevant chapter pages before traveling
- Ensure the annotation tool being demonstrated also supports offline or low-bandwidth use — see Tooling

When distributing printed copies, note the version number and date on the cover. The live site always reflects the latest version.

Collecting Feedback on Usability

After each training session, ask participants:

- Was there anything you did not understand after reading the relevant section?
- Were the examples relevant to your language and context?
- What would you change to make it easier to follow?

Submit responses to the playbook maintainers — the playbook improves through exactly this kind of community use.

Cite this page 2 min read ✦ Contribute

*Last updated on **May 2, 2026** by **Seid Muhie Yimam***

Cost and Resource Planning

Learn how to effectively plan the resources required for dataset creation, including budgeting, timelines, an...

Data Cleaning and Preprocessing

Learn how to prepare raw data for use in language AI systems by improving quality, consistency, and usabil...

Data Provenance and Traceability

Learn how to track the origin, history, and transformations of your data to ensure transparency, reproducibil...

Ethics, Bias, and Governance

Learn how to ensure responsible dataset creation by addressing bias, protecting privacy, and maintaining tr...

Cost and Resource Planning

Learn how to effectively plan the resources required for dataset creation, including budgeting, timelines, and scaling strategies.

Why Planning Matters

Dataset creation can be resource-intensive. Proper planning helps ensure efficient use of time, budget, and human effort while maintaining data quality.

Key Components of Planning

Budgeting Annotation Costs

- **Annotation cost estimation** – Calculate cost per sample or per task
- **Workforce planning** – Consider expert vs crowd annotators
- **Tooling costs** – Include platforms, storage, and infrastructure
- **Quality control costs** – Account for validation and review processes

Time Estimation

- **Task complexity** – More complex tasks require more time per annotation
- **Annotator speed** – Estimate based on pilot studies or benchmarks
- **Project phases** – Include setup, training, annotation, and validation
- **Buffer time** – Plan for delays and iterations

Scaling Strategy

- **Incremental scaling** – Start small and expand gradually
- **Automation support** – Use tools to speed up preprocessing and validation
- **Parallel workflows** – Distribute tasks across multiple annotators
- **Quality vs scale balance** – Maintain data quality while increasing size

*Last updated on **Apr 21, 2026** by **Tadesse Destaw***

Data Cleaning and Preprocessing

Learn how to prepare raw data for use in language AI systems by improving quality, consistency, and usability.

Why Data Cleaning Matters

Raw data often contains noise, inconsistencies, and errors. Proper cleaning and preprocessing ensure that datasets are reliable, accurate, and suitable for downstream tasks such as training and evaluation.

Key Steps in Data Cleaning

Deduplication, Normalization, and Filtering

- **Deduplication** – Remove duplicate entries to avoid bias and overrepresentation
- **Normalization** – Standardize text (e.g., casing, punctuation, encoding)
- **Filtering** – Remove irrelevant, low-quality, or out-of-scope data

Language Detection and Formatting

- **Language detection** – Identify and verify the language of each data instance
- **Formatting** – Ensure consistent structure (e.g., JSON, CSV, text fields)
- **Encoding consistency** – Maintain uniform character encoding (e.g., UTF-8)

Noise and Toxicity Handling

- **Noise removal** – Clean unwanted artifacts such as HTML tags, emojis (if not needed), or corrupted text
- **Toxicity handling** – Detect and manage harmful, offensive, or unsafe content depending on project goals

Missing and Corrupted Data Handling

- **Missing data** – Identify incomplete entries and decide whether to fill, ignore, or remove them
- **Corrupted data** – Detect broken or unreadable content and clean or discard it
- **Validation checks** – Ensure data integrity after preprocessing

1 min read

*Last updated on **Apr 21, 2026** by **Tadesse Destaw***

Data Provenance and Traceability

Learn how to track the origin, history, and transformations of your data to ensure transparency, reproducibility, and accountability.

Why Provenance Matters

Understanding where data comes from and how it has been processed is essential for building trustworthy datasets. Provenance supports reproducibility, enables auditing, and helps identify potential issues in data quality and bias.

Key Components of Provenance

Source Tracking

- **URLs and references** – Record links or original sources of the data
- **Contributors** – Track who collected, created, or provided the data
- **Collection context** – Document when, where, and how the data was obtained

Data Lineage

- **Data evolution** – Track how data changes over time
- **Versioning** – Maintain different versions of datasets
- **Pipeline tracking** – Document each stage of data processing

Transformation Logs

- **Preprocessing steps** – Record cleaning, normalization, and filtering operations
- **Annotation processes** – Track labeling methods and guidelines used
- **Modifications** – Log any changes made to the data after collection
- **Audit trails** – Maintain records for reproducibility and verification

*Last updated on **Apr 21, 2026** by **Tadesse Destaw***

Ethics, Bias, and Governance

Learn how to ensure responsible dataset creation by addressing bias, protecting privacy, and maintaining transparency throughout the data lifecycle.

Why Ethics and Governance Matter

Datasets directly influence the behavior of language AI systems. Poor handling of bias, privacy, or sensitive content can lead to harmful outcomes. Ethical practices and governance frameworks help ensure fairness, accountability, and trust.

Key Components of Ethical Data Practices

Bias Identification and Mitigation

- **Bias detection** – Identify imbalances or skewed representations in data
- **Source bias** – Assess biases introduced by data sources
- **Annotation bias** – Monitor inconsistencies across annotators
- **Mitigation strategies** – Apply re-sampling, re-weighting, or guideline refinement

PII Detection and Removal

- **Personal data identification** – Detect names, addresses, contact details, and identifiers
- **Automated detection tools** – Use models or rules to flag sensitive information
- **Manual review** – Validate automated detection with human checks
- **Data removal or masking** – حذف or obfuscate personal identifiers

Anonymization Strategies

- **De-identification** – Remove or replace identifiable information
- **Pseudonymization** – Substitute identifiers with artificial labels
- **Aggregation** – Present data in grouped form to prevent re-identification
- **Risk assessment** – Evaluate re-identification risks after anonymization

Sensitive Attribute and Content Handling

- **Sensitive attributes** – Gender, ethnicity, religion, health, or political views
- **Content moderation** – Handle harmful, offensive, or explicit content carefully
- **Access control** – Restrict sensitive data to authorized users
- **Use-case alignment** – Decide inclusion based on task requirements

Fair Representation

- **Inclusive sampling** – Ensure diverse representation across groups
- **Balanced datasets** – Avoid over- or under-representation
- **Context awareness** – Consider cultural and linguistic diversity
- **Evaluation fairness** – Test models across different subgroups

Risk Documentation and Transparency

- **Risk identification** – Document potential harms and limitations
- **Datasheets and documentation** – Provide clear dataset descriptions
- **Transparency practices** – Share collection, processing, and annotation details
- **Governance policies** – Define rules for dataset usage and distribution

*Last updated on **Apr 21, 2026** by **Tadesse Destaw***

Annotation Task Design and Human Factors

Task complexity estimation

Inclusive and Bias-Aware Annotation

Learn how to design annotation processes that are inclusive, fair, and aware of potential biases introduced ...

Training and Guidelines

Learn how to prepare annotators through clear instructions, structured training, and iterative refinement of ...

Workflow and Adjudication

Multi-annotator setup

Annotation Task Design and Human Factors

Task complexity estimation

Annotator fatigue

UI/UX considerations

1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Inclusive and Bias-Aware Annotation

Learn how to design annotation processes that are inclusive, fair, and aware of potential biases introduced by annotators and data.

Why Inclusion and Bias Awareness Matter

Annotation is a human-driven process, and annotator backgrounds can influence labeling decisions. Ensuring diversity and awareness helps reduce bias and improves the quality and fairness of datasets.

Key Components

Gender, Age, and Cultural Diversity

- Include annotators from diverse **gender**, **age groups**, and **cultural backgrounds**
- Ensure representation across different **dialects and communities**
- Avoid over-reliance on a single demographic group
- Consider cultural context when interpreting data

Recording Annotator Personas

- Document annotator characteristics where appropriate and ethical
- Capture information such as **language background**, **region**, or **expertise**
- Use anonymized metadata to analyze potential annotation biases
- Ensure privacy and consent when collecting annotator information

Bias Awareness Training

- Train annotators to recognize **personal and cultural biases**
- Provide examples of biased vs unbiased annotations

- Encourage consistent application of guidelines
- Reinforce neutrality and objectivity in labeling

Community Participation

- Engage local communities in the annotation process
- Incorporate native speaker knowledge and cultural insights
- Promote participatory and inclusive dataset creation
- Respect community norms and values throughout the process

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Training and Guidelines

Learn how to prepare annotators through clear instructions, structured training, and iterative refinement of annotation tasks.

Annotation Guidelines with Examples

Clear and consistent guidelines are the foundation of reliable annotation.

- Define each label or category in simple and unambiguous terms
- Provide **positive examples** for each label
- Provide **negative examples** to clarify boundaries
- Include **edge cases** to handle ambiguity
- Specify how to treat uncertain or mixed cases
- Use consistent formatting and terminology throughout the guideline

Training and Calibration Rounds

Training ensures annotators understand and apply guidelines consistently.

- Conduct initial training sessions before annotation begins
- Use calibration tasks to align annotator understanding
- Compare annotations across multiple annotators for the same samples
- Provide structured feedback to resolve misunderstandings
- Repeat calibration until acceptable agreement is reached

Pilot Annotation and Iteration

Pilot annotation helps test and refine the annotation design before scaling.

- Start with a small subset of data
- Identify unclear instructions or confusing labels
- Measure annotation consistency and difficulty

- Collect annotator feedback on task clarity
- Iteratively refine guidelines, labels, and workflow

Minimum Viable Dataset per Task

A minimum viable dataset ensures the task design is valid before full-scale annotation.

- Create a small but representative dataset for each task
- Validate label schema coverage and clarity
- Test annotation workflow and tool usability
- Check feasibility of large-scale annotation
- Use results to decide whether to scale or redesign the task

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Workflow and Adjudication

Multi-annotator setup

Task assignment and redundancy

Disagreement resolution and expert adjudication

1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Data Quality Management

Class imbalance handling

Data Quality Assurance

Inter-Annotator Agreement (Kappa, Alpha)

Data Quality Management

Class imbalance handling

Outlier and noise detection

Error analysis pipelines

Dataset versioning and updates

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Data Quality Assurance

Inter-Annotator Agreement (Kappa, Alpha)

Gold data and control checks

Auditing and spot-checking

Feedback loops

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Image Data

- Image Classification and Recognition – (object classification, scene recognition)

Multimodal Data

- Vision-Language Tasks – (image-text retrieval, captioning, VQA)

Speech Data

- ASR (Automatic Speech Recognition) – (transcription, multilingual ASR, code-switching)

Text Data

- Text Classification – (sentiment, emotion, hate speech, topic, intent detection)

Image Data

- **Image Classification and Recognition** – (object classification, scene recognition)
- **Object Detection and Segmentation** – (bounding boxes, instance/semantic segmentation)
- **Image Captioning and Generation** – (captioning, image-to-text generation)
- **Vision-Language Tasks** – (Visual Question Answering (VQA), referring expressions)
- **Image-to-Image Tasks** – (style transfer, super-resolution, restoration)
- **Document Understanding** – (OCR, layout analysis, form understanding)

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Multimodal Data

- **Vision-Language Tasks** – (image-text retrieval, captioning, VQA)
- **Audio-Text Tasks** – (speech translation, audio captioning)
- **Cross-Modal Retrieval and Alignment** – (text-to-image, image-to-text search)
- **Multimodal Generation** – (text-to-image, text-to-video, image-conditioned text generation)
- **Reasoning and Instruction Tasks** – (multimodal QA, instruction following)

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Speech Data

- **ASR (Automatic Speech Recognition)** – (transcription, multilingual ASR, code-switching)
- **TTS (Text-to-Speech)** – (single-speaker, multi-speaker, expressive TTS)
- **Speech-to-Speech Translation (STS)** – (direct speech translation across languages)
- **Audio Understanding** – (audio classification, sound event detection)
- **Speech emotion recognition**
- **Speaker diarization**

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Text Data

- **Text Classification** – (sentiment, emotion, hate speech, topic, intent detection)
- **Sequence Labeling** – (NER, POS tagging, chunking, slot filling, keyphrase extraction)
- **Sequence-to-Sequence** – (machine translation, summarization, paraphrasing, simplification)
- **Question Answering and Reasoning** – (extractive QA, generative QA, reading comprehension)
- **Retrieval and Ranking** – (document retrieval, semantic search, reranking)
- **Dialogue and Generation** – (chatbots, instruction following, story generation)
- **Structured Prediction and Parsing** – (dependency parsing, constituency parsing, semantic parsing)

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Templates & Artifacts

This page lists all governance toolkit templates available in the playbook. Each template is a standalone doc...

Consent Form Template

When to Use This Template

Contributor Agreement Template

When to Use This Template

Data Ownership Documentation Template

When to Use This Template

Licensing and Compliance

Common License Types for NLP Datasets

Sustainability Plan Template

When to Use This Template

Documentation and Reporting

- Datasheets for datasets

Tooling and Infrastructure

- Annotation tools (usability, scalability, cost)

Data Storage and Release Infrastructure

- Repository hosting (Hugging Face, GitHub, institutional)

Templates & Artifacts

This page lists all governance toolkit templates available in the playbook. Each template is a standalone document you can download, adapt, and use directly in your project.

Template	Who uses it	When
Consent Form	Project leads, coordinators	Before data collection begins
Contributor Agreement	Project leads	When onboarding annotators or chapter authors
Data Ownership Documentation	Project leads, legal reviewers	At project setup and dataset release
Sustainability Plan	Project leads, funders	At project design and end-of-grant stage
Licensing Agreement	Project leads, legal reviewers	Before dataset publication

Other Artifacts

- Annotation guideline template → see Training and Guidelines
- Datasheet template → see Documentation and Reporting
- Data collection plan template → see Cost and Resource Planning

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Consent Form Template

When to Use This Template

What This Form Covers

Part 1 — Project Information

Part 2 — What You Are Agreeing To

Part 3 — Data Use and Storage

Part 4 — Your Rights

Part 5 — Signature

Adapting for Low-Literacy Contexts

Adapting for Oral or Audio Consent

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Contributor Agreement Template

When to Use This Template

What This Agreement Covers

Part 1 — Contributor and Project Details

Part 2 — Scope of Contribution

Part 3 — Intellectual Property and Licensing

Part 4 — Compensation and Attribution

Part 5 — Confidentiality and Data Handling

Part 6 — Termination

Part 7 — Signature

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Data Ownership Documentation Template

When to Use This Template

What This Document Covers

Part 1 — Dataset Identification

Part 2 — Ownership and Rights Holders

Part 3 — Source Data and Third-Party Rights

Part 4 — Contributor Contributions

Part 5 — Licensing and Permitted Uses

Part 6 — Restrictions and Obligations

Part 7 — Contact and Dispute Resolution

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Licensing and Compliance

Common License Types for NLP Datasets

Choosing the Right License

Restrictions and Attribution Requirements

Legal and Ethical Compliance

Licensing Agreement Template

Part 1 — Dataset and Licensor Details

Part 2 — Grant of Rights

Part 3 — Restrictions

Part 4 — Attribution Requirements

Part 5 — Warranty and Liability

Part 6 — Termination

Part 7 — Signature

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Sustainability Plan Template

When to Use This Template

What This Plan Covers

Part 1 — Project and Dataset Overview

Part 2 — Stewardship and Governance After the Grant

Part 3 — Hosting and Access

Part 4 — Community Maintenance and Update Process

Part 5 — Translation and Localization Pipeline

Part 6 — Funding and Resource Strategy Beyond the Grant

Part 7 — Risk and Contingency

Part 8 — Milestones and Review Schedule

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Documentation and Reporting

- Datasheets for datasets
- Standard reporting and reproducibility
- Failure cases and limitations
- Transparency in dataset creation

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Tooling and Infrastructure

- Annotation tools (usability, scalability, cost)
- Data pipelines and automation
- Deployment (cloud vs local)
- Security and access control

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Data Storage and Release Infrastructure

- Repository hosting (Hugging Face, GitHub, institutional)
- File formats and metadata standards
- Versioning and changelogs

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Synthetic Data Creation

- Data augmentation - (paraphrasing, back-translation)

LLM-Assisted Annotation

- LLM-assisted annotation - (human-in-the-loop)

Synthetic Data Creation

- **Data augmentation** - (paraphrasing, back-translation)
- **Fully synthetic generation** - (LLMs, simulation)
- **Scenario-based data generation**
- **Validation against real-world distributions**
- **Evaluation of synthetic data quality**

1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

LLM-Assisted Annotation

- **LLM-assisted annotation** - (human-in-the-loop)
- **LLM and a human annotator agreement**
- **Prompt design and output validation**
- **Bias, hallucination, and consistency control**
- **When NOT to use LLMs**
- **Cost vs quality comparison** (LLM vs human)

1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Data Integrity and Contamination Control

- Preventing train-test leakage

Evaluation, Benchmarking, and Data Integrity

- Evaluation Metrics by task

Data Integrity and Contamination Control

- Preventing train-test leakage
- Overlap with existing benchmarks
- LLM contamination (training data exposure)

 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Evaluation, Benchmarking, and Data Integrity

- Evaluation Metrics by task
- Train/dev/test splits
- Cross-lingual and domain generalization
- Bias and robustness evaluation
- Bias evaluation metrics

Model Building and Starter Kits

- Baseline models for each modality/task
- Training and evaluation scripts
- Reproducibility guidelines
- Benchmark leaderboards
- Benchmark positioning

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Maintenance and Post-Release Strategy

Dataset updates and versioning policy

Release Checklist

Data cleaned and validated

Maintenance and Post-Release Strategy

Dataset updates and versioning policy

Deprecation strategy

Community feedback loops

Issue tracking

 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Release Checklist

Data cleaned and validated

Annotation quality verified

Documentation completed

Licensing defined

Ethical review conducted

Baselines and splits provided

Public access ensured

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Collaboration and Shared Tasks

Shared tasks and benchmarks

Community Ecosystems

Community initiatives (Masakhane, EthioNLP, HausaNLP)

Collaboration and Shared Tasks

Shared tasks and benchmarks

Workshops and open challenges

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Community Ecosystems

Community initiatives (Masakhane, EthioNLP, HausaNLP)

Academic and industry collaboration

Contribution and contributor guidelines

Cite this page 1 min read ✦ Contribute

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Glossary

A reference of terms used throughout the Playbook. Cross-references point back to the chapters where each concept is introduced in depth.

This is a starting point — additions and corrections welcome via the "Edit this page" link at the bottom.

A

Adjudication. The process of resolving disagreements between annotators, typically by a senior annotator or a designated adjudicator. Common when multiple annotators label the same item and a final "gold" label is needed. See the *Annotation Design and Workforce Management* chapter.

Annotation. Attaching structured information — labels, spans, categories, ratings — to raw data so it can be used to train or evaluate language models.

Annotation guidelines. The written specification that tells annotators exactly how to label each kind of input. Includes definitions, decision rules, worked examples, and edge cases. The single most important artifact for high inter-annotator agreement.

Annotation schema. The structural definition of what can be labeled — e.g., the set of allowed entity types in NER, or the rating scale in sentiment analysis. The schema constrains what guidelines can describe.

B

Backtranslation. Translating from the target language back to the source language to generate additional training pairs. Often used to augment low-resource translation datasets. Quality varies — verify with native speakers before training on backtranslated data.

Benchmark. A standardised dataset and evaluation protocol used to compare models. Examples relevant to African NLP: AfriSenti, NaijaSenti, AfriHate, BRIGHTER, AmhEn.

C

Cohen's kappa (κ). An inter-annotator-agreement metric for two annotators on categorical labels, corrected for chance agreement. Range: −1 to 1; conventionally κ > 0.6 is "substantial," κ > 0.8 is "almost perfect."

Consent. Documented permission from the people contributing speech, text, or images, usually including provisions on use, retention, and the right to revoke. Required for ethical and legal data work — see the *Data Collection, Curation, and Governance* chapter.

Corpus *(pl. corpora)*. A structured collection of texts, speech, or other linguistic data used for analysis or model training.

Crowdsourcing. Recruiting many distributed annotators — often online — to label data. Trade-off: scale vs. quality. Quality control techniques (gold-standard items, agreement metrics, qualification tests) become more important as crowd size grows.

D

Dataset. A curated collection of items with labels and documentation, ready to be used for training or evaluation. A dataset is a *corpus + schema + labels + documentation + license*.

Data sovereignty. The principle that data about a community belongs to that community, with associated control over storage, access, and use. Especially important for language data from indigenous and minoritised speakers.

F

Fleiss' kappa. Inter-annotator-agreement metric for more than two annotators on categorical labels — a generalisation of Cohen's kappa.

G

Gold standard. A reference labeling considered correct after adjudication or expert review. Used to evaluate annotators, evaluate models, and as the ground truth in test sets.

I

Inter-annotator agreement (IAA). A quantitative measure of how consistently different annotators produce the same labels. Low IAA suggests guidelines are unclear, the task is ambiguous, or annotators need more training.

K

Krippendorff's alpha (α). A flexible inter-annotator-agreement metric that handles missing data, multiple annotators, and different label scales (nominal, ordinal, interval, ratio).

L

License. The legal terms under which a dataset or piece of code can be used, modified, and redistributed. Common open licenses: Apache 2.0, MIT, CC-BY-SA, CC-BY-NC. Consent and license are different things — covered in the *Documentation, Data Release, and Governance* chapter.

Low-resource language. A language for which little digital data and few NLP resources exist. Most African languages fall in this category. Building useful systems requires deliberate data collection and often careful transfer from related higher-resource languages.

M

Modality. The type of input data — text, speech, image, video, or some combination. Modality-specific annotation is covered in the *Modality-Specific Task Design* chapter.

Multilingual. Covering or working across multiple languages, often with shared model parameters.

N

Named Entity Recognition (NER). Identifying spans of text that refer to named things — people, places, organisations, etc. — and labeling them with their type.

P

Parallel corpus. A corpus with the same content in two or more languages, sentence-aligned. The basis for machine-translation training.

Part-of-speech (POS) tagging. Labeling each token with its grammatical role (noun, verb, adjective, etc.).

R

Reproducibility. The property that another researcher, given the dataset, code, and reported configuration, can re-run the experiment and obtain the same results. The Playbook treats reproducibility as a first-class design goal.

S

Synthetic data. Data generated by a model rather than collected from human sources. Useful for augmentation; risky without verification because errors compound. Covered in the *LLM-Assisted and Synthetic Data Generation* chapter.

T

Tokenisation. Splitting text into the basic units a model operates on. Choices around tokenisation (subword, BPE, SentencePiece, character) materially affect downstream performance — especially in morphologically rich languages.

See also

- How to cite the Playbook
- How to contribute a chapter
- Discord community

*Last updated on **May 1, 2026** by **Shamsudddeen Hassan Muhammad***