



archiving the present



a community guide to
archiving languages,
memories & knowledges

archiving the present

a community guide to
archiving languages,
memories & knowledges

BUKKAFE

... open
speaks

© creative
commons

pāri

Archiving the Present

A Community Guide to Archiving Languages, Memories, and Knowledges

PUBLISHED BY

Co-published by Bukkafe, Creative Commons, OpenSpeaks, and People's Archive of Rural India (PARI).

EDITED BY

Arky AR, Puthiya Purayil Sneha, and Subhashish Panigrahi

COPYRIGHT & LICENSE

Text © 2026 the contributing authors and O Foundation (OpenSpeaks). Cover and interior illustrations © 2026 Siddhesh Gautam. This work, including its illustrations, is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0). To view a copy of this license, visit creativecommons.org/licenses/by-sa/4.0/.

ISBN

978-81-982289-0-1

PRODUCTION

First edition, 50 copies. Released September 2026 at the Archiving the Present workshop and the official Day Zero event of WikiConference India 2026, marking Creative Commons' 25th anniversary.

Cover illustration & design: Siddhesh Gautam

Cover letterpress printing: House of Press, Mumbai, on 450 GSM cotton cardstock

Interior typesetting: OpenSpeaks, set in Scribus 1.7.3, using Source Serif 4, Plus Jakarta Sans, Inter, and IBM Plex Mono

Interior printing: Kolorkode, Bengaluru, on 90 GSM NS Maplitho paper

WITH THE SUPPORT OF

Creative Commons · Wikimedia Foundation · Samagata Foundation

IN PARTNERSHIP WITH

Adivasi Lives Matter · Endangered Languages Archive · Endangered Languages Project · FOSS United · Open Knowledge Initiatives (OKI-IIITH) · People's Archive of Rural India (PARI) · Rising Voices · Whose Knowledge?

The editors introduce themselves after Gratitude; module authors introduce themselves in their own chapter's field notes.

meta.wikimedia.org/wiki/OpenSpeaks

Contents

Preface	5
Introduction	11
1. Planning before documenting	15
<i>Ganesh Birua, Ramjit Tudu and Shalini A</i>	
2. Consent, ethics and payment	27
<i>Taukeer Alam, Anju Gagrai, Opino Gomango, and Puthiya Purayil Sneha</i>	
3. How to record clear audio and video	37
<i>Pius Akumbu and Amrit Sufi</i>	
4. Managing and editing your media files	47
<i>Deepak Kumar Beshra, Amrit Sufi and Subhashish Panigrahi</i>	
5. How to prepare media and metadata for archiving	57
<i>Yulha Lhawa, Abu Sayeed and Subhashish Panigrahi</i>	
6. How to publish your language media	69
<i>Eddie Avila, Mithilesh Baraily, and Mariana Fossatti</i>	
7. Archiving the present: language, storytelling, and languages that live	83
<i>Smita Khator, Ritu Sharma and Sanjib Chaudhary</i>	
8. Raising funds for your language documentation & archiving project	101
<i>Mandana Seyfeddinipur</i>	
A Note on Language	113
Appendix A: Text Style Guide	117
Appendix B: Oral Knowledge Framework	125

Preface

"People are Knowledge" — this title of the 2011 documentary by Achal Prabhala, Priya Sen and Zen Marie stayed with us. Indigenous societies have always kept, protected, and passed on knowledge. Dominant communities have displaced Indigenous peoples, often using state and corporate brute force, and this has broken those knowledge systems apart. Most Indigenous communities have kept their knowledge alive by speaking it, not by writing it down, and that does not make their knowledges any less than written knowledge.

Dominant communities built writing systems, then decided that only written knowledge counts as real knowledge. That took away the dignity of communities that had always preserved their voices and memories through speech. It forced many to hide or deny their own identity.

We also know that technology cannot lift such a structural barrier to linguistic (language-based) and epistemic (knowledge-based) justice. However, it can support new ways for communities to organise. As Corporate AI models expose how information and power are hoarded and abused, more people are seeing the value of building community strength in response. Community documenters and archivists are turning to audio and video to record the voices and living history at risk. On the one hand, it is expensive to record, process and keep audio and video. On the other, audio and video remove access barriers for many—the elderly, those who are illiterate or less literate, and those who just speak a language without a script. Ganesh Birua's work recording Ho songs, stories, and everyday vocabulary is one clear example of what that resistance looks like in practice.

OpenSpeaks has been building open educational resources for community-based language documenters since 2017. We needed these resources ourselves, and often struggled to find them: practical knowledge about documenting a language well was

scattered across different people and projects, not written down anywhere in one place.

This manual brings some of that knowledge together, in the words of the people who hold it. It grew out of a three-month online workshop series, running from 18 May to 24 August 2026 and closing with an in-person session in Kochi, Kerala. Nearly 20 trainers from seven countries across Asia, Africa, Europe, and Latin America led bilingual sessions, each lasting nearly two hours, on their respective parts of the documentation and archiving process. We built each module from the outline a trainer prepared beforehand, then filled it in from the transcript of what they actually said in session: it is their knowledge, in their own voices. Around 104 documenters applied to take part, wanting to document well over a hundred languages between them, including many Adivasi and other tribal languages of India. This book is a work of collective labour: what each trainer actually taught, the question and answer sessions that brought out experiences of the participants and not just the trainers, and the feedback participants shared after each session.

Gratitude

We owe our deepest gratitude to our online workshop trainers, who are also the authors of this book, and the workshop participants. They understood what their communities needed and built their sessions around it, one step at a time.

We also thank our collaborators, who brainstormed with us, shared the workshop invitation widely, and in several cases joined us as trainers: Adivasi Lives Matter, People's Archive of Rural India, Open Knowledge Initiatives, Whose Knowledge?, Rising Voices, the Endangered Languages Project (ELP), the Endangered Languages Archive (ELAR), and FOSS United.

Over the last several years, we have had the privilege of working with communities who speak Bangani, Gorum, Ho, Jaunpuri, Jaunsari, Johari, Juang, Juray, Lambadi, Marcha, Raji, Saptariya Tharu, Sora, and Sri Lankan Malay. This work has been guided by our OpenSpeaks Fellows: Kimmi Pal, Arun Gour, Opino Gomango, Nenavath Mohan, Sanjib Chaudhary, and Uday Raj Aaley, and language collaborators like Omar Rajarathnam (Factum) and Sajehan Buckman. They trusted us with their challenges, openly and vulnerably, and those conversations helped us deepen our focus on real issues.

That trust also shaped how we ran the workshop. We asked applicants who showed real promise to sign up as trainers, not only as learners. Of the 104 people who applied, 37 signed up to train fellow participants too.

Ansh Arora of FOSS United has stayed with us in the background for months, and Sai Rahul Poruri, also of FOSS United, has supported us throughout. Ranjith S helped build Subtitler, the tool that has subtitled countless videos from this workshop series.

TinkerHub and TinkerSpace Kochi lent us space and ran a hardware support campaign; Jasim CM and Mehar MP made that happen. They also connected us with Afeef, who introduced us to Muhsin C of Bukkafe, our publisher.

We are grateful to Jacqueline Chen of the Wikimedia Foundation, who has guided OpenSpeaks' strategic direction with real generosity; to our advisors Mandana Seyfeddinipur, Padmini Ray Murray, Chinmayee Mishra, and Jnanaranjan Sahu; and to the South Asia Regional Funds Committee, who engaged with our work for months and made it better.

Ritu Sharma and Smita Khator of PARI have advised us from the start of this workshop, well beyond their own module. Their

module gave this book its title. We could not find a better way to say what this book sets out to do than to keep collective stories and memories alive.

This book would not exist without Siddhesh Gautam's cover art, whose anti-caste activism has left a lasting impression on millions, including us. Murtaza Sethwala and Mohammad of House of Press turned that art into a letterpress print, and Kolorkode is printing the book itself. Every cover is hand-pressed, leaving a physical impression on the paper. We hope this book leaves a similar one: helping you record your community's stories, and helping those stories leave their long-lasting mark.

Finally, we thank the Wikimedia Foundation, the Samagata Foundation and Roshni Murali in particular, and Creative Commons, for their support. And we thank our editors: Arky AR, who said yes without hesitating over the extra work it meant, and Puthiya Purayil Sneha, a trainer in this workshop series as well as an author and editor of this book.

Arky AR, Puthiya Purayil Sneha and Subhashish Panigrahi

15 August 2026

About the editors

Arky AR is a technologist and visual storyteller with over a decade of experience building Free and Open Source communities and collaborating with various non-profits across Asia, Southeast Asia, and Africa, including Braille Without Borders (BWB), NGO Resource Center, Mozilla, and Allen Institute for Artificial Intelligence (AI2).

Puthiya Purayil Sneha is a researcher with the Open Knowledge Initiatives programme, International Institute of Information Technology, Hyderabad, India. Her areas of interest and work

include digital media and culture, open knowledge, digital rights and gender justice.

Subhashish Panigrahi is a language documenter, archivist, and documentary filmmaker, and the founder of OpenSpeaks. He has led community, language and justice and rights programmes across Asia-Pacific for Ashoka, the Wikimedia Foundation, Internet Society and Mozilla.

Introduction

We created *Archiving the Present* as a practical guide about documenting and archiving community stories, voices and knowledge, in mother tongues and local languages. It covers planning a documentation project, recording audio and video, managing files, preparing metadata, publishing, archiving, fundraising, and the ethics of consent and payment.

It exists because oral knowledge is still treated as inferior to written knowledge. Most human knowledge has always been oral. Community documenters correct that by recording everyday stories and traditional cultural memories, and by building archives that communities themselves control and manage.

This manual is written for community documenters. You may be working alone. Most communities have only a handful of documenters or archivists, and few have regular access to someone running a training session, so there is often no trainer or workshop to guide you. Each module teaches its ideas directly to you, the way a written guide teaches a method to someone reading it alone, not the way a facilitator's script teaches a room. This manual was first planned as a guide for community-based trainers, but that framing did not hold up: most community documenters do not think of themselves as trainers. Within a community, the same person often learns, teaches, and reviews others' work. A hard line between "trainer" and "documenter" does not match how that work actually happens. If you do want to use this manual to teach someone else, one-to-one or in a workshop, each module ends with a section called "Teaching this module", gathering what you need for that.

The manual has eight modules, each covering one part of the documentation process, followed by two appendices: a text style guide for captioning, subtitling, and transcribing, and the Oral Knowledge Framework on which this manual's practice is built.

Within each module, the main text teaches the idea itself. Alongside it, boxes marked "Field notes" tell a real story: a documenter's own learning from lived experience, a specific archive, a specific community. The field notes are not summaries of the main text; they show what the ideas above look like when someone has actually done them.

This manual is deliberately written in simple English. Most of its authors and potential readers speak English as a third or fourth language. Not being fluent in English should never be the reason someone cannot learn from this book.

You do not need to read this manual from start to finish before you begin. Already know what you want to document? Start with Module 1 and work through the parts that apply to your project. Looking for an answer to one specific problem, such as how to record clear audio or how to write fair consent into your process? Go directly to that module.

We are keeping a digital version of this manual online, and will keep it updated as practice changes. Your community may translate it by machine. The plain, direct English used throughout this manual is written to survive that translation without losing its meaning.

Begin documenting now, with whatever equipment and time you have, rather than waiting until you can reach a perceived gold standard of documentation. The modules in this manual will help you do that well.

1. Planning before documenting

Ganesh Birua, Ramjit Tudu and Shalini A

This module helps you plan your language documentation around your own community's motivations. By the end, you should be able to write a simple plan: what is missing and needs documenting, who your community collaborators are, and which stories, words, or songs you will record. The plan should also say why this work matters, what its goal is, and where the recordings will live safely for the long term.

1.1 What is important but not recorded

Ganesh Birua

Think about your own village, town, or community. What topics have not been recorded yet? A ritual, a festival, farming or forest work, a local issue, a craft, women's work, migration; each is a starting point. Notice who the knowledge keepers are: elders and others who hold important stories, songs, or history. Notice the power dynamics; who holds more power in the relationship between you, the documenter, and the person you interview? And notice memory places: a local river, a shrine, a meeting place, a market, or a forest path. The main question behind all of this is simple: "What do we know is important, but have not yet recorded in audio or video?"

We often see only what is right in front of us. But tribal/Indigenous communities have protected their knowledge, traditions, and way of life through oral storytelling for many generations. Today, much of this knowledge is slowly disappearing. Traditional seed preservation, forest food collection, childcare practices, traditional midwives (*dhai*), village priests (*deuri*), and even community identity are changing due to migration, displacement, and outside influences. If these stories and practices go unrecorded now, future generations may never know them.

Not having a writing system (script) for your own language is not

a reason to avoid documenting it. You can still write it down using another script; this is called transliteration. The absence of a native script should never stop you from documenting your language.

Field notes:

Ganesh Birua is a language activist and archivist from Mayurbhanj district, Odisha, and has documented his native language, Ho, for about six years. Since 2017, he has created hundreds of educational images, run the Warang Chiti Project 2030, and collaborated with OpenSpeaks Archives on transcriptions and translations.

He sees a concrete example of a documentation gap: in his area, Particularly Vulnerable Tribal Groups (PVTGs) such as the Juang and Mankidia live nearby. Many young people from these communities no longer know their own community's folk songs, stories, or rituals, having moved away for education or work.

Everyday vocabulary is also at risk. Words for birds, plants, and other things found in the forest can disappear if no one writes them down. Ganesh has used only a Word file to record local bird names in Ho because very few of these names appear elsewhere in writing.

Rituals themselves are changing. The Ho community traditionally marks three life events with distinct rituals: birth, marriage, and death. Even the birth ritual, including the practice of naming a child after an ancestor, is slowly shifting under outside influence. "Document these details now, while elders still remember them, because the details themselves are disappearing fast," says Ganesh.

1.2 Who to interview and how

Ganesh Birua

If you are documenting songs, stories, or traditions, ask around your community to find out who is widely known for their knowledge of these. Listen carefully during public meetings, festivals, or everyday conversations, too, since these often reveal who explains things clearly and thoughtfully.

Approach interviewees in two steps. First, make contact by visiting, calling, or messaging to introduce yourself and explain your interest. Then hold a proper meeting to explain your project, your partners, and what will happen to the recording. It is perfectly fine if someone says no or needs time to think. Good planning always gives people time and choice.

Field notes:

Ganesh's documentation gap example in §1.1 offers guidance here, too. In his experience, elders and people who still live in the village often hold the most knowledge, while youth who have moved away for school or work often no longer have it. This is a practical reason to prioritise reaching out to elders and long-term residents early, rather than delaying.

If you visit a village for an interview, first make people feel comfortable. Let them know you also come from a village and understand their way of life; this helps them feel you are one of them, not an outsider. Village people value love, respect, and friendly relationships, so do not start with serious questions. Begin with topics such as their festivals, songs, dances, farming, or village life. When people feel relaxed, they share their stories and knowledge more willingly.

1.3 Plan for both visits and "on-the-go" recording

Ramjit Tudu

Language documentation happens in two equally valid ways. Planned visits are where you fix a specific time and place with the speaker and prepare beforehand. On-the-go recordings are where you record spontaneous moments while working in the field, attending a festival, or travelling.

Keep your phone or recorder charged, with some free disk space, so you are ready when the moment comes. Make a quick note of the date, place, and names immediately after any sudden recording. Think of these on-the-go recordings as part of your plan, not as mistakes. Consent still matters even when a recording is spontaneous; §1.5 covers what that looks like in practice.

Keep in mind: if you see or hear something important and safe to record, it is better to record quickly, then follow up later with a calmer, planned conversation.

1.4 Decide where the recordings will live first

Ramjit Tudu

Your recordings need a home, and there are different kinds to choose from: local community spaces, specialised platforms, and large online archives. Community spaces include local groups, libraries, youth clubs, schools, community centres, and local collectives. Platforms include community-focused organisations such as People's Archive of Rural India (PARI)¹, large online platforms such as the Wikimedia Commons (<https://commons.wikimedia.org>), and national repositories.

Community spaces come first, because your primary responsibility belongs to the people you record. Add external archives and large platforms later, once you have secured clear consent and people feel completely comfortable.

Field notes:

Ramjit Tudu is a Santali language activist and technologist from Mayurbhanj, Odisha, India. He co-founded Ol Chiki Tech, created the Birmali e-magazine, and is a key contributor to Santali Wikipedia.²

His own priority order for a recording's first home would be: community spaces like local groups, local libraries, youth clubs, schools, or community centres, since the goal is for a recording to reach the people it actually belongs to.

Alongside that, archives like People's Archive of Rural India (PARI) can also hold recordings; collaborating with an organisation that works closely with communities lets them help host the material.

Large online platforms come after that: Wikimedia projects, like Wikimedia Commons, and other archives can host recordings so they reach a wider public, once licensing requirements are checked. Hosting there also makes it easy for anyone to find and access the material later online.

The “home” you choose matters for another reason: always note down metadata, the place a ritual, song, or custom belongs to, since the same subject can look different from one place to another. Recording when and where something happens, not just what happens, is what lets people later understand the recording correctly.

1.5 Connect planning with consent, safety, and labour

Shalini A

Connect your documentation plan to three simple questions.

Consent: "How will I explain clearly what I am doing, and where the recording may go?" Safety: "Could this recording put anyone

at risk if it is shared, and how can I reduce that risk?" Labour:

"Who is giving their time and knowledge, who is doing the physical work, and how will we compensate them?"

Consent: Consent is an ongoing process, not a one-time "yes". It must be voluntary, based on clear advance information, and specific to the current recording. Speakers can also withdraw their consent at any time by telling the documenter. Consent can be a simple gesture ('implied consent'), a clear spoken agreement ("express consent"), or a signed document ("informed consent"). For most community documentation, a clear spoken agreement is enough. Always explain when you are recording, how you will store it, and where it might go before you start recording.

Remember that the knowledge holder owns the recording, not the documenter. Always respect requests to withdraw knowledge already shared, even if someone else has recorded it in the past.

Safety: Sharing private information carelessly can put people at risk of harassment, public shame, or professional trouble. Badly stored data can be hacked, and people can manipulate or clip recordings out of context to spread misinformation.

To reduce this risk, share only what is necessary and avoid releasing additional private details. Know who will actually see or use what is published. Keep recordings on secure, private platforms rather than posting everything openly on the internet. Mask names, locations, or voices when someone does not want to

be identified. Delete unneeded material, including backup copies, once it is no longer needed.

Watermarks can also help prevent unauthorised redistribution. Under Indian law, Article 21 protects the Right to Privacy, which gives people legal recourse if others violate their privacy.

Labour: Documentation requires work from many people, not just the documenter. It includes community members who share their knowledge, elders and knowledge keepers who pass down traditions, and researchers, institutions, or regional NGOs who support the process. Recognise this labour by respecting that the community owns its own knowledge and data. Always credit specific individuals and communities by name, rather than using an anonymous group label. Finally, compensate people fairly for their work. This can mean direct payment, investing in community capacity, or hiring community members as paid researchers and co-authors, rather than treating them as mere research subjects.

Field notes:

Shalini A is a community archivist and technologist at Janastu and Servalots. She also leads community Wi-Fi mesh and low-literacy projects, and created the BaaApp and Follow the Sheep initiatives with shepherd communities. At Janastu, she leads many community archiving initiatives focusing on women's issues and caste-based stories across the Indian state of Karnataka, with work also extending into the state of Tamil Nadu.

Consent: Shalini's practical addition to the consent checklist above is the "show-back." After recording a photo, audio, or video, show the person what you captured and ask again whether they're comfortable with it being kept or published. Consent given before recording isn't the end of the conversation; confirming it again once they have seen or heard

the actual result is what makes it real.

Labour: Documentation labour includes the documenter's own time and effort, not only the time and effort of the person being interviewed. Both deserve recognition. Budget for accommodation costs alongside travel where relevant, and treat the person and the place being documented with genuine respect; that matters as much as getting the technical process right.

Ramjit Tudu adds three further points from his own documentation practice:

On consent: Some tribal communities have their own internal restrictions on certain knowledge. They may want particular knowledge to stay within the community and not be disclosed outside it. Respect that boundary, and only share what people have clearly agreed to share, nothing beyond it.

On licensing: Consent conversations should also cover how the material will be licensed. Some contributors are comfortable releasing their material under an open licence; others prefer more restricted terms. Publish according to whatever the contributor has actually agreed to, not a default choice made for them.

On labour, payment and a travel budget: Documenting someone nearby is manageable, but reaching a subject farther away costs more, and that cost should be planned for, including through fundraising if needed, rather than absorbed informally. Where a performer or knowledge-holder is directly involved in a recording, budget to pay them for that time and expertise, too; for example, someone demonstrating a song, craft, or ritual.

Before your first few recordings, write down who will be the speaker (or speakers), who will accompany or help you during the recording, and any worries you already see, for example, sensitive topics, borders, land conflicts, or gender-based risks.

These notes will be useful again in Module 2, when you learn specific ways to record consent, decide access levels, and handle payments.

1.6 Teaching this module

If you only document and archive a language, you can skip this section: it is for teaching others.

Session goal: Help the people you're teaching build a simple, written plan for their own language documentation, grounded in their own community's motivations, not an outside template. By the end, each person should be able to name what is missing and needs documenting, who their community collaborators are, and which stories, words, or songs they will record.

A few concrete ways to run it:

For §1.1, open by asking the group, or the person you're mentoring, to think about their own village, town, or community, and list what has not yet been recorded: topics, knowledge keepers, power dynamics, and memory places. This works well as a shared exercise before you introduce the field notes.

For §1.5, close the session by asking each person to write down who will be the speaker, who will accompany or help them, and any worries they already have. This is the same planning exercise the module ends with for a solo reader, and it works just as well done together.

Across the module, use the field notes as your material. They are written to be told as stories.

-
1. <https://ruralindiaonline.org>
 2. <https://sat.wikipedia.org>

Notes

Notes

2. Consent, ethics and payment

Taukeer Alam, Anju Gagrai, Opino Gomango, and Puthiya Purayil Sneha

This module explains how informed consent, the rights of participants and documenters, FAIR (Findable, Accessible, Interoperable, Reusable) and CARE (Collective Benefit, Authority to Control, Responsibility, Ethics)-based ethical practice (see Appendix B), and payment work together in community language documentation. By the end, you should be able to ask for consent properly, respect community and individual rights, and plan for fair payment and recognition of labour.

2.1 Asking for consent and recording it

Taukeer Alam

Consent rests on three pillars: respect, trust and protection. Respect means valuing the speaker's time, knowledge, and comfort: always ask, never assume. Trust means building a relationship before you turn on the recorder; trust is the foundation of documentation. Protection means making sure the recording does not harm the speaker or the community.

Consent means more than a yes. The speaker needs to know why you're documenting; whether it is for community preservation, research, or something you plan to share more widely. Where the recording will live matters just as much: a local archive, Wikimedia Commons, the Endangered Languages Archive (ELAR)¹, or a commercial platform. So does how it gets used later: in teaching, on public websites, or across Wikimedia projects.

Levels of access to the documentation matter too. At the fully public end, anyone on the internet can see or hear the recording (for example, recordings on Wikimedia Commons). Restricted access narrows that down to researchers or specific community members. Private or community-only access keeps the recording inside the community altogether, never to be shared online.

Consent also has levels. Personal or individual consent comes

from the person being recorded. Family or household consent is needed when you record inside a home, show family members, or share family-specific histories. Community consent is needed when documenting community-owned knowledge, sacred rituals, or traditions; in many villages, this means speaking with the village leaders or elders first.

In many oral communities, a written consent form can feel formal, or in some cases, raise outright suspicion. An oral consent statement, recorded on audio or video, works just as well. A short, local-language script helps here: have the speaker state their name, date, language, and what they agree to share, before you start recording.

Respect requests for privacy and anonymity. Some speakers may want only audio to be recorded, or ask to keep their name and identifying details off the record. Be ready to stop or delete, too. If a speaker or community member changes their mind during the conversation or later, respect their decision to withdraw consent, pause, or delete the recording where they ask.

Field notes:

Understand local risks as well. Forest-dwelling communities such as the Van Gujjars may show wood or fire usage while documenting traditional knowledge. For example, how they make mawa (also khoya), a dairy product made from reduced milk, in alpine meadows; this can draw negative attention from forest departments or other authorities. So always discuss the risks of sharing such recordings publicly before you do. Sacred or ritual knowledge carries its own risk: recording a traditional healer's chants, for instance, might be seen by other villagers as practising witchcraft, which can put both the healer and the documenter in danger. If a conflict like this arises, apologise, respect the community's rules, pay any fines imposed by the community, and delete the recording in front of the

community. Safety comes first.

Field notes:

Opino Gomango, a linguist and Sora-language activist has worked with the Living Tongues Institute for fifteen years, documenting language and culture in the Sora-Jurai language cluster. On one visit to a village in Odisha, he met a shaman from his own Sora community and, after discussing it together, began recording the shaman's practice. Partway through, some villagers heard about the recording and objected; they said the shaman was known for using traditional healing practices for the wrong reasons, and asked Opino to stop. He stopped the recording, apologised, and explained that the recording was his own idea, made for research, and not the shaman's fault. Once people understood what the research was for, the situation settled.

Even with consent from the person being recorded, a wider community's concerns can still surface mid-recording. Being ready to stop, apologise, and explain as part of the plan, not as a failure of it.

2.2 Rights, licenses, and ownership

Anju Gagrai

Traditional knowledge holders have both moral and legal rights to their stories. Documenters must obtain explicit permission before they record, transcribe, or archive sacred ceremonies and histories.

Ownership works differently for different people. The community holds cultural ownership: the folktales, songs, recipes, and cultural knowledge itself belong to the community, the speaker as its master maker. The documenter or researcher holds the copyright of the physical or digital recording, but must

share that right with the community as co-ownership. Simply transcribing or writing down a community's stories never gives a researcher, university, or author sole ownership of that heritage or the right to tell it however they like. That is extraction, not documentation.

A licence is a digital permission slip attached to the file, telling others how they can reuse it. Wikimedia Commons, for instance, uses open Creative Commons² licences that let people share, copy, and reuse the files, even with changes. Sensitive material needs more caution: get the community's agreement on the licence, pick one that rules out commercial use, or simply keep the material off public platforms.

2.3 Payment and labour

Puthiya Purayil Sneha

Sharing traditional knowledge is work. Recalling stories from memory, translating, and transcribing take real time and energy, both mental and emotional. This labour is not shared evenly: women, for instance, often carry significant work to keep traditions and domestic rituals alive, and their contribution can go unrecognised. Documenters must recognise every contributor: speakers, translators, transcribers, community reviewers, and anyone who lends a space for the documentation to happen.

Compensate speakers for their time, travel expenses, and any work they miss, especially daily wage earners. Keep payments modest and reasonable; a large sum can pressure people to say what they think the researcher wants to hear, which is its own kind of harm. Remember what you are paying for: the labour, time, and trouble of sharing, not the knowledge itself, which is priceless. If cash payment is not possible or appropriate, there

are other ways to compensate: copies of the recordings or photos for the family, food, tea, and transport, public thanks or a mention in a community book, or gift vouchers and other tokens of appreciation.³

If a speaker asks to remain anonymous for safety, remember that digital payments (like UPI in India) or a bank transfer leave a paper trail. Use cash or gift vouchers instead to protect their identity. Always take the community into confidence before you begin documentation, and ask them directly what form of compensation is appropriate. For a worked example of fair, equitable pay across roles, including how to account for compounding marginalisation, see Appendix B.

2.4 What this looks like in practice

Puthiya Purayil Sneha

Many researchers from outside the community collect data and publish books without giving credit or fair payment. This is extraction, and it costs trust; some community members have even destroyed their own materials rather than see them extracted again. Treat community members as co-creators, not research subjects.

Standard consent and payment guidelines, the kind institutional review boards use, often fail in rural and tribal contexts. Consent has to be situated in the community's own realities: border conflicts, electricity shortages, gender dynamics, and whatever else actually shapes daily life there.

Field notes:

Opino Gomango has seen what extraction looks like up close. A Sora community artist and researcher, he has spent over twenty-five years, since 1997, writing down Sora grammar, songs, and stories by hand. A researcher from outside the

community visited him, took his manuscripts with the promise of publishing them, and paid him a small amount of money. The book was published, but it did not carrye artist's name, not even an acknowledgement. The only trace of him was an uncaptioned photo of him and his wife on the cover. The extraction caused deep trauma. He burned his remaining manuscripts and decided he would rather let his knowledge die with him than write or speak about Sora culture again.

Extraction destroys trust and hurts communities. Documenters must always give full credit, respect co-ownership, and ensure fair recognition.

Keeping a language alive is a heavy emotional burden. Document that burden itself as part of your metadata: who holds the knowledge, and why it is difficult for them to preserve it.

If a recording puts anyone at risk, whether from witchcraft accusations, political vulnerability, or local conflict, etc., safety comes first, always. Be willing to stop, apologise, and delete files to protect the community, even if it means losing the recording.

2.5 Teaching this module

If you only document and archive a language, you can skip this section: it is for teaching others.

Session goal: help the people you're teaching understand how consent, rights, ethical practice, and payment work together in community language documentation, so they can ask for consent properly, respect community and individual rights, and plan for fair payment.

A few concrete ways to run it:

For §2.1, run the Consent and Access Circle activity. Draw four concentric circles on a board or paper: Level 1 (inner), private,

kept within the family or community only; Level 2, restricted, stored in an archive with controlled access for researchers; Level 3, public, shared online but not widely promoted; Level 4 (outer), fully public, openly shared on Wikimedia Commons, websites, and social media. Give the group four scenarios to place on the circles: a grandmother singing a children's lullaby; a traditional healer chanting a sacred healing prayer; a village elder sharing a historical story about a land conflict; a young person singing a modern song in their mother tongue. For each one, ask which circle it belongs to, what type of consent (oral, written, family, community) it needs, and how the contributors should be credited or compensated. Close by discussing how consent can change over time; for example, a family later asking to remove a public recording.

Across the module, use the field notes as your material.

-
1. <https://www.elararchive.org>
 2. <https://creativecommons.org/cc-licenses/>
 3. https://meta.wikimedia.org/wiki/OpenSpeaks/Oral_Knowledge_Framework#Remuneration

Notes

Notes

3. How to record clear audio and video

Pius Akumbu and Amrit Sufi

This module aims to help you record clear, honest audio and video with the tools you already have, in ways that respect people's comfort and safety.

3.1 Prepare people and space

Pius Akumbu

Greet each speaker warmly and explain the recording clearly before you start. Walk them through what to expect: the format, how long it will take, and what you will ask them to do. Respecting their time by being upfront builds trust from the first minute. Listen to their voice early, too: a loud or soft speaker changes how you should set up your microphone.

Pick a quiet room for the best sound quality, and watch out for hidden background noise: a humming refrigerator can quietly ruin an otherwise good recording. Leave enough space for speakers to move, gesture, or even stand up and dance; these physical movements often carry meaning of their own. Let anyone nearby know you are recording, so they do not walk in and interrupt.

Provide comfortable seating, food, water, and childcare if speakers need it. Let them know they can pause or stop at any time; taking a break is completely fine.

Field notes:

Pius Akumbu is a Cameroonian linguist and one of the language revitalisation mentors at the Endangered Languages Project (ELP). In one of his recordings in Cameroon, a refrigerator standing behind his speaker position hummed quietly enough to miss by ear, but the microphone picked it up clearly, and it nearly spoiled the recording. He checks for exactly this kind of background sound before starting now. He

also keeps his camera framing wide enough to capture a speaker's full range of movement: in one of his recordings, the people he was speaking with broke into song and started dancing partway through, and because his frame was wide, the moment was captured rather than cut off.

3.2 Use phones and other devices well

Pius Akumbu

Phones, audio recorders, and video cameras can all produce quality recordings when used well. Phones usually have good settings out of the box, but check the settings on dedicated recorders and cameras carefully. If you use an external microphone, which is recommended, verify every connection and setting before you start, and always record a test clip and listen back to make sure the audio actually works.

Set up your recording device between you and the speaker, close enough to capture their voice clearly, and put on headphones to monitor the sound live. If you must hold your phone, keep it as steady as you can and move slowly when following any action; a tripod gives a more stable video. Test your lighting before you start: it affects video quality just as much as background noise affects audio.

Start every recording with a full battery, and check that your SD card or phone storage has plenty of free space. Transfer earlier recordings to an external hard drive regularly, rather than letting files pile up on one card.

Field notes:

For group recordings, Pius recommends one cardioid microphone placed centrally, rather than passing a single microphone between speakers. In one of his own recordings, a cardioid mic on the table was enough to clearly capture five

speakers sitting around it; passing a microphone from person to person slows a conversation down and makes people wait their turn to speak.

Equipment can also fail mid-recording: one of Pius's own recordings stopped abruptly when his camera overheated. DSLR-style cameras often have a maximum continuous recording time, sometimes as short as 20 minutes, before they need a short break to cool down; check your own device's limit in advance, and plan for short pauses if you are using this kind of camera.

On resolution: full HD (1080p) is usually enough for archival purposes. Higher settings, 4K and above, fill storage cards much faster, so calculate the recording time you need against your card's storage before you start; free calculators are available online.

3.3 Audio/Sound: clear voice

Pius Akumbu

If you are using a lavalier (clip-on) microphone, keep it about 15–30cm (6–12 inches) from the speaker's mouth, close enough for a clear voice, far enough that clothing does not rub against it. Control loud background sounds like traffic, animals, or loudspeakers, but keep some ambient sound to preserve the local atmosphere; background noise should add to the scene without drowning out the speaker's voice. Always record a short test clip and listen to it together with the speaker before the real recording starts.

Field notes:

A soft-spoken interviewee (Pius originally used “consultant” for interviewed speakers) should never be asked to speak louder for the sake of the recording; respect their natural

volume. Instead, move the microphone closer or use your device's gain setting to slightly boost the recorded volume. Gain has a cost, though: turning it up also picks up more background noise, so use it carefully and only as much as needed.

3.4 Picture: frame, light, and movement

Amrit Sufi

Frame shots with dignity: never from above, never cutting off the head or body, and free of distracting background clutter. This is not about looking professional. It is about letting the speaker move naturally without disappearing from the shot. Use natural light whenever possible, with the speaker facing the light source rather than having it behind them. Leave extra space in the frame so the speaker stays in shot when they gesture or shift in their seat. Agree on simple silent hand cues beforehand (a raised hand for "you're done," a thumbs-up for "all good") so you can communicate without interrupting the speaker. Keep any side conversations quiet if someone is helping you record, so the speaker feels like the centre of attention.

Field notes:

Amrit Sufi is an Indian language activist and speaker of Angika, a language spoken across Bihar and Jharkhand in India, and parts of Nepal. She founded the Indic Oral Culture project, which supports South Asian citizen archivists in documenting their own languages online. In some of her own recordings, framing has gone wrong: an interviewee's head cut off at the top of the frame in one, and in another clip, an interviewee sliding sideways until only half their body remained in view. Her fix is to set the frame wider than feels necessary at the start, so there is space for people to move naturally without being cut off.

3.5 Emotional safety while recording

Amrit Sufi

Never push anyone to share painful memories, no matter how valuable the information feels to your project. If someone becomes upset, pause or stop recording and let them decide when or if they want to continue. Sharing your own stories, when it feels appropriate, builds a sense of equality between you and your speaker and helps them feel more comfortable opening up. Documentation work can be emotional for the documenter, too. This is not a weakness; it is part of the work. Prepare yourself mentally, and know your own limits.

Field notes:

While recording a relative singing a *vidaai*, a farewell song traditionally sung when a daughter leaves her parental home after marriage, Amrit's relative became emotional, remembering her own daughter. Amrit paused the recording immediately, gave her relative the time she needed without pressure, and, because they already knew each other well, was able to check in personally rather than treat it as just another interview.

A strongly personal or political opinion is not, on its own, a reason to stop recording. Check afterwards, without the camera on, whether the speaker is comfortable with what they said being kept. If you are unsure whether something can be published, you can still record it for the archive with the interviewee's consent and decide together later whether and how it gets shared.

In group settings, some voices, often the more powerful or privileged ones, can crowd out others. Notice this as part of your fieldwork, not just an inconvenience.

Field notes:

Journalist Smita Khator uses a practical trick from her own reporting work: inviting the women of a household to a private moment, for example, by asking to be shown to the washroom, to ask what they could not say in front of the men of the family.

Where shame or social stigma keeps someone, often a woman, from appearing on camera, take time to understand the reasons before pushing forward. Explain clearly why you are recording and where it will go, and reassure them that they keep control, including the right to have a recording deleted.

Oral history is rarely a simple set of facts. People share stories, songs, opinions, and things they have heard from others. When something said on record seems disputed, the right response is usually not to fact-check it on the spot, but to record it clearly as that person's own account, and note this in your metadata.

3.6 File formats and tools

MKV (Matroska) for video and WAV for audio are future-proof, archive-preferred formats that are easy to convert between. MOV is acceptable but not ideal, and some archives accept MP4 for video.

HandBrake (free, open source, available at <https://handbrake.fr/>) converts video and audio between formats. Audacity (free, open source, available at <https://www.audacityteam.org/>) removes background noise: select a noisy stretch of audio, "train" it on that noise, then remove similar noise from the rest of the file. This also works on audio extracted from a video, which can then be cleaned and synced back.

For more on what a finished, publishable piece of documentation

can look like, PARI's own submission guidelines are a useful reference.¹

3.7 Teaching this module

If you only document and archive a language, you can skip this section: it is for teaching others.

Session goal: Help a group record clear, honest audio and video with the tools they already have, in ways that respect people's comfort and safety.

A few concrete ways to run it:

Work in pairs to record a 3–5 minute practice interview in a local language.

Watch or listen back to the footage together and identify one thing that worked well and one thing to improve across three areas: sound, picture, and the speaker's comfort.

Ask each other, "Would you be happy if this were shared widely?" and discuss why or why not: a quick way to practise checking for consent.

1. PARI (People's Archive of Rural India) submission guidelines, <https://ruralindiaonline.org/submission-guidelines>

Notes

Notes

4. Managing and editing your media files

Deepak Kumar Beshra, Amrit Sufi and Subhashish Panigrahi

This module is to help you organise your audio and video simply, keep them safe, and do basic editing so your recordings are ready for sharing, archiving, and future use.

4.1 Naming folders and files

Deepak Kumar Beshra

Clear folder and file names save time later, when you come back to edit, upload, or archive your material. A simple folder structure works well, for example:

Collection/

00_admin/ (consent notes, project notes)

01_raw/ (original recordings)

02_edit/ (edited versions)

03_transcripts/ (text, subtitles)

04_access/ (compressed copies for sharing)

05_uploads/ (files prepared for Wikimedia or other platforms)

Add a few key details to each file name: date, language, a person's name if that is safe to include, place, and topic, so a name means something without opening the file. Try this yourself: take two or three of your own file names and rewrite each as a "before" (the messy original) and an "after" (a clear version); seeing the difference side by side makes the value obvious.

Field notes:

Deepak Kumar Beshra is an Indian filmmaker and the founder of the Santali Art and Film Development Foundation. While editing the Santali-language films he directs, he creates one main folder with five subfolders, always in the same order:

video, raw audio (recorded separately with an external mic), photographs, project/editing files, and the final exported output. Footage stays traceable months later this way.

His file-naming convention is `description_PersonName_date`. An interview recorded with Devendra Singh on 21 December, for example, becomes `interview_DevendraSingh_21Dec`: the description first, then the person, then the date, always in that order. He never leaves a file with only a camera's default numbering, like `IMG_0042.png`. Without a proper name, even he can't tell what's in a file without opening it.

When several photographs or clips look almost identical, common with burst-shot photography, he keeps all the raw files in one folder and copies in another. In the copy folder, he keeps only the ones he'll actually use, in a separate "edited/selected" folder. He then deletes the unused raw duplicates if storage is tight and the edited copies are safe.

4.2 Basic editing: from raw file to useful copy

Deepak Kumar Beshra

Two versions of each file matter: the "master," the original, high-quality version you keep safe and untouched, and the "access copy," a smaller version optimised for phones, easy sharing, or quick uploads.

Always work on a duplicate, never on the original master file: editing your only copy risks losing it for good. Trim long empty silences at the start and end, and boost the volume gently if a recording is too quiet, without reaching for heavy filters that distort the natural sound. A recording does not need to be perfect: the goal is clarity and understanding, not cinematic production.

Field notes:

Deepak's first rule in editing: never touch a raw file directly. Always duplicate it first and do all editing on the copy, keeping the original safe in its own separate folder in case something goes wrong.

This is what that discipline looks like on a real project, a documentary about how rubber is collected and processed. In his edit, he imported the video and a separately recorded audio track, synced the audio to the video, trimmed to the useful parts, and added B-roll (a shot of someone collecting rubber sap) alongside the main interview footage to show visually what was being described. He also records a few extra seconds before a person starts speaking, so the very beginning of what they say is never lost.

OpenCut, a newer free and open-source video editor (an open-source alternative to CapCut), is worth trying if you want a capable editor that is also free to use (<https://opencut.app/>).

4.3 Backups and simple tracking

Amrit Sufi

Losing a file can mean losing a unique song, story, or voice that can never be recorded again. Backing up is not extra work you do if there's time; it is what stands between a recording and losing it for good. Keep at least two copies of every important recording in separate places: your laptop and an external drive, for example, or shared between two trusted people. Where you can, distribute files across group or community storage rather than letting everything depend on a single person.

A simple tracking log helps you keep sight of your own progress: a list or spreadsheet with columns for file name, date, speaker

code(s), topic or genre, and status (raw, edited, subtitled, uploaded, deposited).

Field notes:

Amrit Sufi's classmate once documented Gojri, a nomadic language spoken in Jammu, India, recording strangers he met on buses and while walking, their folk songs and conversations. He lost all of it when his phone broke, since none of it had been backed up anywhere else, and could never find the same people again, since he had met them only by chance, in a particular place, at a particular time. A person's memory of the same story changes on a second telling; the original moment, once lost, cannot really be recovered.

Her own spreadsheet tracks files with columns for category (folk song, proverb, interview, B-roll, and so on), a description, the month and year recorded, and status (transcribed, digitised, uploaded, and where). With everything in one table, she can see at a glance what's done and what's still pending, without replaying every file in her collection.

4.4 Transcripts and subtitles

Subhashish Panigrahi

Even a short piece of text adds real value to a recording: it helps people who cannot listen to audio easily, makes material easier to reuse for teaching and online platforms, and helps archives and future listeners understand what they are looking at.

A full transcript is not always realistic, and it does not need to be the goal. A short summary in one or two languages, a transcript of just one important story or song from a longer recording, or a simple subtitle file for one short clip are all realistic starting points. Whatever you produce, note the names of people, place, date, and a brief line on what is happening: this context matters

as much as the exact words spoken. For the full captioning, subtitling, and transcript formatting conventions, see Appendix A.

Field notes:

Subhashish Panigrahi conceptualised and built early prototypes of OpenSpeaks Subtitler (<https://subtitler.toolforge.org/>), a subtitling tool built by Ranjith S. It helps a user: load a video, mark where a segment of speech starts and ends, listen and type what was said into the matching box, then move to the next segment. The result exports as SRT, an open standard supported by YouTube and most other video players and platforms. OpenSpeaks Subtitler will support an automatic speech-to-text feature in the future, giving time codes and subtitle text that might be slightly inaccurate, needing a human editor to correct them.

4.5 Teaching this module

If you only document and archive a language, you can skip this section: it is for teaching others.

Session goal: Help a group organise their audio and video simply, keep it safe, and do basic editing so recordings are ready for sharing, archiving, and future use.

A few concrete ways to run it:

Ask two or three people to describe one of their current file names, and work together to suggest a clearer, more organised one.

Play a very short audio or video clip in its raw form, then a lightly edited version, and name the three changes made, for example, trimming the start and end, raising the overall volume, or removing a sensitive sentence.

Share a brief, real-world example of a file that was lost or recovered. If you have one, open the floor for questions on backups.

Ask one person to describe a recorded story or song in a few lines, then show the group how those lines could become a short description or the start of a transcript.

Notes

Notes

Notes

5. How to prepare media and metadata for archiving

Yulha Lhawa, Abu Sayeed and Subhashish Panigrahi

This module is to help you understand what media and metadata are, and prepare both so that archives, libraries, and other memory institutions can work with the material you collect. This builds on Module 4, which covered getting files into good shape. Archives do not expect perfection; they expect clarity: clear file names, clear descriptions, and clear access rules.

5.1 What archives need

Yulha Lhawa

Connect what you already have (files, notes) to what archives usually ask for.

Every archive needs two things from you. Your "media" is the actual files you have collected: audio, video, text, and images, including field notes you scribble down when you forget your recorder, and photographs you take alongside a recording. Your "metadata" is information about who is speaking, where, when, in what language, and what the recording is about.

Most communities have only one or two people actively documenting their language. If something happens to that person's files, and those files were never archived, that work is lost for good — nobody else can ever access it. Archiving is what turns a personal recording into something that a community and future generations can actually use.

This does not mean your recording has to be perfect. A slightly noisy file from an elder in their 80s or 90s is still worth archiving because it may be the only recording of that story that will ever exist. What matters more than perfect audio quality is a clear file name, a clear description, and a clear note on who can access it and how.

Field notes:

Yulha Lhawa is a Tibetan linguist who works for the Endangered Languages Project (ELP), a US-based nonprofit, where she mentors community documenters from around the world. She grew up documenting her own mother tongue, Khroskyabs, a minoritised and endangered language of the Tibetan Plateau spoken by only around 3,000 people, and has been doing language documentation and revitalisation work since high school. Yulha treats metadata almost like part of the recording itself. At the beginning of an audio file, she records a short spoken note that includes who is recording, who is being recorded, what the story is about, the date, and consent, so she never has to reconstruct these details from memory later.

5.2 Media: choosing and organising archive-ready files

Yulha Lhawa

Decide which files to keep as "masters" (final versions), which to share as "access copies," and how to group them for deposit.

Keep two versions of each important recording: a "master" and an "access copy." The master is the best-quality, untouched original; never edit or convert this file directly, and keep it safely to one side. The access copy is a lighter version for phones, sharing, and everyday use, like a photocopy of the master, so if anything goes wrong while editing or sharing, the original master is never at risk. WAV for audio and MP4 for video are the standard formats, since both are widely supported by almost any device or software and are comparatively smaller files. Formats like MOV take up far more space for the same recording (even three or four MOV files can fill a drive), so if your camera records in MOV format, keep the original instead of repeatedly

converting it, which reduces quality. For images, JPEG and TIFF are the two common choices.

Keep file names short: a language code (find yours in the ISO 639 language code list¹ if your language has a two- or three-letter code), a number, and maybe a person or date. Avoid putting long, multilingual descriptions in the file name itself, since software used for transcription and archiving, such as ELAN or FLE_x, can struggle with long file names and you may end up renaming your whole collection later; save fuller, multilingual detail for the folder name and the metadata instead. If you separate languages in a folder name, use something consistent, like a colon or a forward slash, the way the Endangered Languages Archive (ELAR) does automatically. There is no fixed rule about which language should come first: some documenters prioritise their own community's language, rather than defaulting to English or another dominant language, to show that people need not know English or Hindi to use the archive.

Write dates as year-month-date, for example 20260713, in file names and in metadata, since different countries write dates differently and year-month-date is the closest thing to a shared convention in language documentation. Keep everything in one master project folder rather than spreading files across your phone, laptop, and other places. Decide on a logic for organising sub-folders: by speaker, by date, or something else. Write that choice down in your metadata, since even you are likely to forget the reason behind it later. Keep folder structures shallow, three or four levels deep at most, so nobody has to click through a dozen nested folders to find one file. For each small project, one storyteller, one event, one topic, create a folder containing the final master files, the access copies, and a short text file or sheet listing what is inside: that list is the beginning of your metadata, covered next.

Field notes:

Yulha used to keep long file names, as she wanted every detail packed in so she would not have to open a file to know what was inside. That changed once she started using ELAN (a transcription tool) and FLEEx (a dictionary-building tool), both of which struggle with long file names, forcing her to rename her entire collection. Her rule of thumb now: keep file names well under ten characters where possible. She recommends browsing ELAR's own online collections before archiving anywhere yourself: some are very well organised, some are messy, and seeing both makes it easier to decide what to aim for.

5.3 Metadata: describing each file clearly

Abu Sayeed

Show you the minimum information to collect for each file so archives and communities can understand and reuse it.

Focus on a small, core set of fields. This is not about writing everything down, but about writing down enough.

For each file:

Title: a short, meaningful name.

Language(s): which language(s) are spoken.

People: speaker(s), recorder, interviewer; use codes if names are sensitive.

Place: village/town, region, country.

Date: recording date, or at least year.

Type: story, song, conversation, interview, vocabulary list,

and so on.

Access and licence: whether it is public, restricted, or community-only, and what licence applies, based on your consent decisions.

For the collection or project as a whole, add a short description, three to five lines on why and how the recordings were made, note who is involved (community groups, organisations, archives), and explain how people can request access, if it is not fully open.

How much detail you add depends on your collection and who is likely to use it. A general documentation project can stick to the core fields above. A more specific collection, say of plant and forest names, might need extra detail that a specialist would care about, such as what recording equipment or settings you used. Think about who is likely to use your collection: general community members, linguists, or researchers, and let that guide how much detail you add. A short, complete set of core fields is worth more than a long, half-finished one.

If people in your community are more comfortable in a regional language than in English, consider writing key metadata fields in both languages, for example, English and Hindi, or English and your documentation language. This makes your collection genuinely usable by the people it is about, not just by outside researchers.

Field notes:

Abu Sayeed is a Bangladeshi language activist and keeper of the endangered Sylheti script. He is working to revive and promote the Sylheti Nagri script through Wikimedia projects. Much of his own metadata practice happens through Wikimedia Commons: he uploads photographs, manuscripts, and his own writing, including a sonnet he wrote and published himself,

recording basic details like the date collected and where it was generated, and publishing the work under a free licence so it can be reused for education and other purposes. A useful lesson from community documentation is that metadata should serve both people and archives: it should be understandable to community members while also being structured enough for future researchers, archivists, and digital platforms to use.

5.4 Tools and templates that can help

Abu Sayeed and Subhashish Panigrahi

You can always record metadata without needing complex software.

You do not need special software to begin. A simple metadata sheet, on paper or digital, covering the core fields above, is enough to start. You can enter the same information into a tool like Tome (<https://tome.toolforge.org>) or a plain spreadsheet. Both can export metadata in formats archives commonly use, such as JSON, CSV, or Wikitext, and let you move from a basic sheet to more specialised tools later, as you need to. A filled-out example, even just one file worked through completely, is often more useful to copy from than an empty template.

Field notes:

For translating metadata into other languages before publishing, Abu Sayeed's go-to is Tome (<https://tome.toolforge.org>), a web app that can translate entries, for example, from English into Hindi, so that documentation stays understandable to people who don't read English. His own convention for Wikimedia Commons: keep the file title short so it is easy for others to find and translate, and put the fuller context in the description field. He also uses his own website and blog to publish transcripts and interviews

in other languages, alongside or instead of uploading to Commons.

The Internet Archive (<https://archive.org>) is a fully public archive: only put things there that you are genuinely comfortable with the whole world seeing; anything private or sensitive should go somewhere else, or stay off it entirely. In one of Abu Sayeed's own interviews, a woman told a story and sang a song about her language, but did not want her face or name shown; in the final edited video, her face was blurred, keeping her story and her consented material without exposing what she'd asked to keep private. For quick, offline documentation, his own template is simple: a title, the people involved, the date, where the story came from, and a note on how to use it. The goal is consistent, understandable records that remain useful to the community and to future generations.

On self-hosted archival tools, ones that keep data on your own computer or server rather than a third-party platform, two options worth knowing are: Lameta, a free, open-source tool built specifically for language documentation by a consortium of language archives, which exports metadata in open formats like JSON and CSV, and Murkutu CMS (<https://mukurtu.org>), a community-controlled platform that Yulha Lhawa also recommends. On using transliteration, writing one language's words in another script, such as Tamil words in Latin letters, in metadata: it is not prohibited and is fairly common in language archives, especially for non-Latin scripts, but because transliteration systems are often not standardised, Yulha's own guidance is to avoid leaning on it too heavily, since it can end up confusing rather than clarifying.

5.5 Linking media and metadata to actual archives

Subhashish Panigrahi

Connect the media and metadata you have prepared to an actual

archive, and see what a simple linking system looks like in practice.

Once your files are named and organised, and you have written down metadata for them, you still need a way to connect the files to that information. This step can be simple. You can keep information in an Excel sheet. Many documenters use a text-editing app on their phones or even handwritten notes in a notebook, which they later type into a document after returning home from a recording. An online sheet, like Google Sheets, can be easily shared with others, allowing multiple people to edit it together. But for very sensitive information, private spreadsheets in LibreOffice or Excel work better.

Larger institutional archives use dedicated metadata tools. The Endangered Languages Archive, for instance, recommends Lameta, a free, open-source tool built specifically for language documentation by a group of language archives. It exports your metadata in open formats like JSON and CSV, so you can share it with others in your community. Tools like this take some practice to learn. If you are starting your project, you can begin with paper or a notes app, then move to more advanced tools later.

Archives (community archives, language archives), repositories (Internet Archive, Pad.ma²), platforms (Wikimedia Commons), and other GLAM institutions (galleries, libraries, archives, and museums) will generally ask you for two things: your media files and metadata about them, often as a spreadsheet or database. By following this module, you should be able to select a small set of files, write basic metadata for them, and decide on an access level based on the consent you have collected.

Module 7 goes further into what archiving means for the present, storytelling, and community power. This module has covered the

practical basics needed to even start that conversation.

5.6 Teaching this module

If you only document and archive a language, you can skip this section: it is for teaching others.

Session goal: Help a group understand what media and metadata are, and prepare both so archives, libraries, and other memory institutions can work with what they collect.

A few concrete ways to run it:

Work through one real file together, filling out the core metadata fields (§5.3) completely for it, rather than starting from an empty template — a single worked example is easier to copy from than a blank form.

Try Yulha Lhawa's own habit: record a short spoken note at the start of a file (who is recording, who is being recorded, what the story is about, the date, and consent), then compare notes on what everyone chose to include.

If some in the group are more comfortable in a regional language than in English, practise writing one file's metadata in both languages, and discuss what changes in the translation.

1. <https://www.loc.gov/standards/iso639-2/langhome.html>

2. <https://pad.ma>

Notes

Notes

6. How to publish your language media

Eddie Avila, Mithilesh Baraily, and Mariana Fossatti

This module explains why publishing matters, who should be involved, and simple ways to share language recordings and texts safely and usefully.

6.1 Language documentation as digital activism

Eddie Avila

Documenting a language and sharing it online is itself a form of digital activism. It means communities using the internet and digital platforms to represent themselves on their own terms, including in their own language — whether that means growing a body of text, audio, and video content, teaching the language to others, or helping create new speakers.

This does not require being a professional linguist, academic, or trained documenter. It needs the same protocols already covered earlier in this manual: consulting with the community, deciding together how to share it, and agreeing on who gets access to it.

There is no single "right" platform for sharing content. The choice depends on where a community's audience already is, and how much control the community wants over its own data. Reaching more people quickly (for example, on YouTube) often means less control over licensing and reuse. More open platforms, like Wikimedia Commons, involve different trade-offs around licensing and control. Look at what other communities are already doing, and learn from their successes and failures, rather than searching for one perfect tool.

Rising Voices' own *Digital initiatives for Indigenous languages* toolkit¹ It is one starting point for strategies and real community examples. Specific tools named in resources like this can quickly go out of date, even though the underlying strategies remain useful.

Field notes:

Eddie Avila is the managing director of Global Voices, where he also directs Rising Voices, the organisation's digital inclusion arm. Rising Voices works with communities around the world who want to use the internet and digital platforms to represent themselves on their own terms, including in their own languages. This is what Rising Voices calls "digital activism for languages": people taking concrete action to promote, teach, and preserve a language online, whether that means growing a body of text, audio, and video content, or creating new speakers. Five years ago, Rising Voices partnered with UNESCO during the International Decade of Indigenous Languages to publish a 190-page toolkit of strategies and real examples from communities worldwide. Many of its specific tools are already outdated, as platforms change every few years, but the underlying strategies hold. Choosing a platform or tool is a personal decision that depends on where a community's audience already is, and on how much control the community wants over its own data. For example, YouTube reaches more people quickly, but a more open platform like Wikimedia Commons involves different trade-offs around licensing and control. Rising Voices' own role is less about telling communities which tool to use, and more about connecting people to others doing similar work, so communities can learn from each other's successes and failures.

6.2 Why publish

Mithilesh Baraily

Publication is an important step in documentation. It lets you share stories outside communities in a curated and accessible manner. During the documentation process, consult with community members to understand their interests and

motivations in publishing. Publishing can make the language or community visible beyond its immediate circle. It supports teaching, community pride, and advocacy. It attracts allies, funders, and future collaborators. It can tell multiple stories of the region when a new political, religious, or other development takes place. And it lets people learn about the knowledge and culture of a particular region, such as its folklore, history, herbs, flora, and fauna.

Publishing does not mean "put everything online." It is sharing the right things with the right people; sometimes it means putting everything on social media so interested people will connect with each other, or contact you. That is how you find the people who will actually help protect and carry the documentation forward, based on the consent and access decisions covered in the earlier modules.

Field notes:

Mithilesh Barailly is a Himalayan social entrepreneur and researcher dedicated to documenting the endangered Kami language and preserving the vital linguistic heritage of artisan communities. For Mithilesh, publishing in or about language is not about pursuing personal gain or fame, but about safeguarding the language. He began documenting his own language because he wanted an identity for his community in a world where they faced discrimination based on work and caste, and their traditional occupation. When he first asked people in his community why they should publish their language, they told him there was pride in being a distinct community, and that publishing was itself a form of cultural reformation. He compared publishing to opening a restaurant on the internet: a lot of people will love what you share, but not everyone will take the time to read or learn from it — and that is fine. What matters is that publishing reconnects a scattered community (a diaspora spread across the world) to a shared

sense of belonging and pride, and it can open doors to researchers, funders, and organisations who want to support language preservation. Publishing does not mean putting everything online, though. Mithilesh comes from a family of blacksmiths, and some of his community's traditional techniques and designs are knowledge that is deliberately kept within the community. Deciding what stays private and what gets published has to be guided by the same ethical consent principles covered earlier in this manual.

6.3 Who gets to publish

Mithilesh Baraily

People who can be involved in publishing include community members: speakers, youth, and elders. They can include local organisations, such as schools, collectives, and archives. And they can include external partners like researchers, Wikimedia volunteers, and GLAM (galleries, libraries, archives and museums) staff.

Two questions are worth asking before anything is published: who should decide what is published and what stays private, and who is responsible for checking consent and licences beforehand.

This should be shared decision-making, not one person or one institution deciding alone.

Field notes:

The different people who can be involved in publishing include community members, academic researchers, and external partners such as Wikimedia volunteers or GLAM staff. Specifically, where research is involved, community members should never be treated as research subjects from whom data is extracted. Instead, they should be treated as co-authors or the

principal investigators, who hold the consent and have a real say in what gets published and what does not. Mithilesh shared a hard lesson from his own fieldwork: years ago, he brought community elders forward to support a new cultural and development board, only to watch people from his community use that goodwill for political gain without doing any of the ongoing work to preserve the language. His community's language is spoken on both sides of the India–Nepal border, raising questions of legitimacy across it. His own approach is to think beyond politics and borders altogether — caste, religion, and community do not follow national lines, and cross-border history has shaped the language on both sides. He has also seen a pattern worth watching for when documenting men and women in a community. During one field visit, an elderly woman clearly knew the language and its stories best, having learned them directly from her grandparents, but she kept pushing a younger male relative to speak instead of answering herself during the interview. The lesson: make space for the person who actually holds the knowledge, not just the person who steps forward to speak.

6.4 Where to publish

Mariana Fossatti

Publishing also means deciding where documentation will live. There are two broad spaces to weigh. Community spaces include local meetings, schools, radio, WhatsApp groups, shared folders, and printed booklets. Public and digital spaces include Wikimedia Commons,² Wikidata,³ Wikipedia,⁴ Internet Archive⁵, community websites, blogs, social media, and language archives.

These are options that community members need to consider carefully. Which archives should live in each space? Can everything go public, and what is needed for something to get published, in terms of licences, descriptions, metadata, media

quality, and so on? Who should be able to access the archives? What should be shared first, or only, first, among community members? And how do the archives move between the two spaces: are community members allowed to share them outside communal spaces, is it possible for the general public to ask for permission, get access, and share under certain conditions, like under Creative Commons and other licences?

Pick one community space and one public/digital space that makes sense for your current project. These options are flexible and not mutually exclusive: while not all archives would go public, a specific item can live in community and public spaces at the same time.

Public spaces are excellent for visibility. They are also very hard to control: once content is downloaded, it can be shared far beyond the platform it was published on. Some public digital spaces offer more nuanced options, like Creative Commons licences, which give some level of choice and control back.

Archivo Inundación is one example. It is a project that collects community memories from a catastrophic flood that occurred in 2003 in Santa Fe, Argentina. Affected communities provided photographs, home videos, brochures, alternative and out-of-print publications, personal diaries, and other materials for digitisation. Each contributor then decided whether the digital archives would remain private, be shared on the project website⁶ with some restrictions, or be shared more openly and unrestricted on Wikimedia Commons.

Archives (Module 7) and funding (Module 8) can support publishing and long-term access to documented content.

Field notes:

Mariana Fossatti is a feminist and a free/libre culture activist from Uruguay, and a long-term Wikimedian. She is the

Decolonizing Wikipedia coordinator of Whose Knowledge?, an organisation focused on knowledge justice on the internet, and leads its Decolonising Wikimedia programme. Wikimedia Commons (a sister platform to Wikipedia with more than 100,000 freely usable media files) is one option among the public/digital spaces mentioned in this module for publishing of documented content, alongside the community spaces discussed earlier. Commons only supports a small set of open licences, though, and not every choice a community might want for its own cultural material is available there. Because of this, some communities keep their own website or communal archive first, where they can collect consent from everyone involved, and only later decide what, if anything, moves to Commons. Archivo Inundación (the Flood Archive) makes this concrete. It is a project from Santa Fe, Argentina, led in collaboration between members of the affected communities and Wikimedistas del Litoral User Group (now known as WikiActivistas del Litoral), a group of Wikimedia volunteers.⁷ The city was hit by a severe flood in 2003; twenty years later, local Wikimedians worked with community members to collect photographs, home videos, and other memories of the disaster. Some contributors agreed to share their material only on the project's own website. Others also agreed to upload it to Wikimedia Commons, where the material now lives in its own dedicated category, with detailed documentation so visitors can see exactly where it is from.

6.5 How to publish

Mariana Fossatti

A simple four-step path can apply to most platforms.

- 1. Select material:** choose files that are ready, with clear audio/video and basic metadata, and where consent for sharing has been acquired.

2. Check consent and licence: Confirm that speakers and communities agree to this type of publication, and decide the licence and access level (for example, open, restricted, community-only). For Wikimedia Commons, only two Creative Commons licences are available: CC BY and CC BY-SA. These are the most open licences that allow any use of the files (including commercial use). The difference is that CC BY-SA asks for reciprocity — people who remix the files in their own work are required to share it under the same licence.

3. Prepare description: Write a short description of who, where, when, which language(s), and what the recording is about. Before including any names or identifying details, check whether the speaker or community asked for anonymity earlier in the process (see Module 2, §2.1). Add tags or categories if the platform uses them (for example, language name, region, topic). For Wikimedia Commons, it is advisable to browse the platform to find suitable categories for the material. If new categories need to be created, put them under a higher-order category that already contains them, so the material stays easy to discover.

4. Upload or share: For local sharing, plan the event, group, or format (screening, booklet, radio segment, WhatsApp clip). For online sharing, follow the basic steps of the chosen platform (Commons upload, page creation, or posting), using the prepared files and descriptions.

Uploading to Wikimedia Commons:

The Wikimedia Commons (<https://commons.wikimedia.org/>) lets anyone upload audio, video, image and PDF documents. It is a sister site of Wikipedia. The images or audio/video you see on Wikipedia mostly come from Wikimedia Commons. Commons is a platform used primarily for educational purposes, completely free of charge, ad-free, and supported by

a community of volunteers who carefully maintain the content. It is a good option for hosting content that communities may want to share openly.

You can create an account for free. After logging in, use the “Upload Wizard” to drag and drop photos you have clicked or audio or video files you have recorded. The wizard’s first real check is a declaration of who made the work and whether it is entirely original.

Commons itself has no separate field to record community consent, so this check must be completed before the upload, not during it.

For the licence, you can choose any of the licences you see there. Mariana uses Creative Commons Attribution-ShareAlike (CC BY-SA 4.0) because of the ShareAlike reciprocity clause—anyone who remixes the work has to release their new version under the same licence. This is the most open licence available on Commons that still asks for something in return from re-users, although it does not give the community full control, the way some other licences might.

Then fill in a caption, a longer description, and metadata (structured data) tags that link the file to Wikidata, and describe categories, which are “the shelf of your archive”. The more specific and well-chosen the category, the easier it is for others to find the file. Archivo Inundación used the same pattern, a dedicated category, with sub-categories for each affected village, so a whole collection stays organised and discoverable.

For audio, Commons supports the most common formats: mp3, WAV, OGG (OGA) and FLAC. For video, it supports WebM, OGV, MPEG-1, and MPEG-2, but not MOV or MP4. For files outside Commons’ supported formats, Clip to Commons, a connected tool, can convert and upload files under 20 minutes long in one step.

A Creative Commons licence does not protect community recordings from being scraped to train AI models. CC licences were written before generative AI, and today, content is scraped across the open web, including the Commons. If something must never leave community control, do not put it online at all. Publishing openly and keeping something fully private are the two real options: no licence in between currently guarantees protection against scraping.

6.6 Teaching this module

If you only document and archive a language, you can skip this section: it is for teaching others.

Session goal: Help a group decide, safely and deliberately, where and how to publish their language documentation, so they can weigh platform trade-offs, agree on who publishes what, and choose a community space and a public space that fit their own project.

A few concrete ways to run it:

For §6.4, ask each person or group to pick one community space and one public or digital space that fits their own project, then work through the three questions together: which archives belong in each space, who should be able to access them, and how material moves between the two. Bring your own platform-specific examples if you have them, such as one Commons file, one small community booklet, or one radio story.

For §6.3, open with Mithilesh's two questions, who should decide what gets published and who checks consent and licences, as a short discussion, using his field note as a starting point.

Across the module, use the field notes as your teaching material. Eddie's, Mithilesh's, and Mariana's examples above give the group

real cases to work through.

1. <https://unesdoc.unesco.org/ark:/48223/pf00003871in86>
2. <https://commons.wikimedia.org>
3. <https://www.wikidata.org>
4. <https://www.wikipedia.org/>
5. <https://archive.org/>
6. <https://archivoinundacion.ar/>
7. More about this project: <https://diff.wikimedia.org/2023/12/20/look-mum-the-photo-you-took-is-on-wikipedia>

Notes

Notes

7. Archiving the present: language, storytelling, and languages that live

Smita Khator, Ritu Sharma and Sanjib Chaudhary

This module does not see an archive as a one-time act that preserves the past. Instead, it sees it as an ongoing practice that serves your community in the present. By the end, you should be able to name at least one thing beyond a word list worth documenting in your own community, and explain why the context around a recording matters as much as the recording itself.

7.1 Why archive the present, not just the past

Smita Khator

Most people carry one of two pictures of an archive in their heads. The first is a safe box: somewhere to keep something valuable so it survives after its last speaker is gone. The second is a record of the past, rarely of the present. Both pictures miss the same thing. Preserving a language and sustaining a language are different jobs. Preservation happens once. Sustaining a language never stops, and it needs a community of living people to carry it.

Government bodies, private companies, and researchers already run many language documentation projects. AdiWani is one example. What they often lack is utility for the community itself. A museum-style gaze treats a language as an artefact, cut off from the social and political reality of the people who speak it. Resist that gaze in your own work. Document the present as it actually is. Do not fall into the trap of saving only a pristine, edited-down version of the past. Let us remember, an archive is not just a place where old things are stored. It decides what gets saved, named, displayed, hidden, digitised, and returned.

Before you start recording, ask three questions: who contributes to this documentation, who benefits from it, and what does it do for the present and the future. Do not stop at who might use the archive one day.

There is a stark difference between community documentation of Indigenous repositories and documentation done by non-community individuals or institutions. You can see the gap in the objectives and utility. If you document as an outsider, position yourself as an ally, not the authority. Inside the word "ally", there is a power hierarchy that doesn't disappear just because our intentions are good. We usually have the equipment. Often, the funding relationship. Often, the institutional letterhead. That is a hierarchy, however we call it. When a funder offers resources for preservation, what the community stands to lose and gain is a collective decision, never one person's call to make.

Community archiving survives only if the community sustains it. Train other members as you go, so the practice continues even when you are not there to lead it.

Start documenting now, whatever the funding situation, whatever the material's chances of being called "publishable." What eventually goes out into the world, and what stays inside the community, are separate questions. The community answers them later, on its own terms.

Documentation in the form of a first-person account reads differently from a report written about someone. Both have their place. But the gaze is not the same, and this manual will not pretend otherwise.

Field notes:

Smita Khator is the Chief Translations Editor of PARIBhasha and leads PARI's operations in 16 Indian languages. She has worked in the area of archiving, is a Bangla translator, and also writes on women's issues and labour. Misa17 (Misael Elias), an 18-year-old Ecuadorian rapper, performs in A'ingae (Cofan), a language spoken by around 1,500 people. His songs are about resisting the oil and mining companies that have entered his

community's land, and about not losing his grandparents' language and voice. Nobody records the language as it is actually spoken today, in its mixed, borrowed, changing, imperfect form, because older material always feels more valuable, more archival, more worth funding. The trap is documenting a rescued, pristine version of the past instead of the present. Two real museums show what this looks like in practice. The Madhya Pradesh Tribal Museum, a government museum in Bhopal, describes its mission as capturing tribal life in a modern building. Yet its published advisory team and curators include hardly anyone from the ghumantu (pastoralist) communities whose dictionary the museum is building. Compare this with the Museum of Santali Culture in Bishnubati, Birbhum, West Bengal, built and run entirely by the Santal village itself: the men thatched the roof, the women plastered the walls with mud and cow dung as they do in their own homes, and students donated objects and now look after the collections. It stays open to outside visitors, but does not depend on them. Demati Dei Sabar, known as Salihan, is a Sabar freedom fighter from Odisha. P. Sainath documented her story in his book *The Last Heroes: Foot Soldiers of Indian Freedom*. Her community has lived through cultural appropriation and extraction twice over. Neela Madhaba, her community's own deity, was appropriated as Lord Jagannath by Brahminical powers. Decades later, the manapatra, the only official document proving Salihan was a freedom fighter, was stolen by journalists who had promised her family recognition and support in return. Sainath's own writing names the second theft plainly: one journalist exploiting a Sabar woman for a story, reported on by another journalist who refused to look away. The industry was the same. Who each writer answered to was not.

7.2 Building blocks of an archive

Smita Khator

Archiving borrows two crucial principles from journalism: objectivity and verification. Objectivity is often used interchangeably with neutrality, but they are not the same. Objectivity means recording what is actually there, accurately, including the parts that complicate your own argument. Neutrality means refusing to take a position at all, and an archive of an endangered language cannot be neutral. Making it is already taking a position. What it must be, instead, is objective. Without that, the record is worthless. Verification works the same way. Unverified material does not become true simply by sitting in an archive for thirty years.

Begin documenting now, not once you can reach some gold standard. Start with what you have. Even a basic smartphone is enough. You are not competing with professional archivists working to achieve institutional standards. We are a resource-strapped nation. Indigenous communities are the worst victims of historical inequality. Hence, you start with what you have. The struggle for access to better things must go on, but that should not deter you from documenting your own community's heritage for your children, and theirs.

Field notes:

Archiving is not something colonial institutions invented. Before colonial institutions collected Indigenous knowledge, Indigenous people already had their own archival systems. Memory lived in oral histories, songs, ceremonies, place names, clan systems, and kinship responsibilities, carried in everyday life rather than stored in a building. PARI's language universe positions itself against that history and aims to

decolonise and de-romanticise the idea of the archive.

7.3 Ownership and access: what a digitisation project taught us

Smita Khator

Two real projects show what ownership questions look like in practice.

The first involved digitising centuries-old Bangla *pothi/puthi* (palm-leaf manuscripts). The practical work came first: basic resources, a working light connection, somewhere to sit. Only once that was in place did the harder question surface. Who owns the result? The first digitised copy always went to the funders, such as the British Library, because they had paid for the digitisation. Anyone else who wanted to see it had to go through the British Library.

The second project was a hundred-year-old local library, about to be demolished to make way for a community recreation centre. It turned out to be infested with termites. Cleaning, cataloguing, and physically preserving the books came first. Digitisation is often the last step, not the first.

Both projects raise the same question. Who controls the archive, and on whose terms and conditions does outside funding arrive? In both cases, the material itself became secondary to the funder's conditions. Aim instead to collect and preserve on an ongoing basis, not as a one-time act tied to a single grant.

Today, institutions often say, "We digitised it so everyone can access it."

But digitisation can repeat harm when sacred materials become searchable, ancestral photos circulate without consent, recordings are uploaded to the public domain, or incorrect labels

become database truth. Even a digital archive can still act as a colonial archive to control and perpetuate stereotypes and aim for assimilation.

Many colonial archives grew through conquest, omission, missionisation, surveys, maps, census, boarding schools, anthropology, and government surveillance. Ancestors were removed from studies. Languages were recorded by outsiders.

So, Indigenous data sovereignty means Indigenous Peoples have the right to govern data about their people, lands, cultures, knowledge, and futures. This includes who collects it, stores it, describes it, accesses it, benefits from it, and decides when it should be restricted or returned.

And let us remember, for someone outside the community, archiving might be a passion project or part of a paid job. For a community member, it is their way of life and an integral part of their survival, carried on top of existing work and livelihood, like a Santali museum curator who is a full-time farmer.

Field notes:

Digitising centuries-old cultural documents, administrative papers and other material under the British Library's Endangered Archives Project. Khator's erstwhile institution did the labour but did not get complete ownership: the first digitised copy in formats set as standards by the ELP went to the British Library, since it had funded the digitisation, and anyone in India who wanted to see it either used the British Library's own link or made an appointment to view it on her team's screen. Before any of that, the more basic problem was to create a space for documentation with even the most basic of requirements, such as electricity. Phase one of these projects was always arranging light, tables, chairs, and extension cords. These least romantic parts of archiving were actually the most of it.

A second story: a hundred-year-old library in Bihar's Patna was about to be demolished to make way for a community recreation centre. Her team had seven days. The books, including Bangla grammars and primers dated 1852, sat in cartons in a damp, dark loft, heavily termite-infested, pages torn and spines broken. What they did first was not glamorous: catalogue, clean, and stabilise. Physical preservation comes first. Digitisation is what you do once something has survived.

One real case, from an earlier cohort of this training: a community member's life's work documenting his own community, built over two decades, was co-opted by an outside scholar who published it without ever crediting him. Khator's advice to community members going into a funding relationship: voice a collective position, not an individual one, since a community speaking as one is harder to exploit than a person negotiating alone. Before signing anything, write down plainly what is gained and what is lost. A funder's resources always come with something asked in return, whether that is credit, the first copy, or the right to decide who sees the material.

7.4 What makes an archive: data and contextual metadata

Ritu Sharma

Language documentation has traditional tools: word lists such as the Swadesh list or the Anvita Abbi list, pictorial elicitation books such as the Frog Story, and primers or grammars. These tools are useful, but documenting a language's lexicon is not the same as preserving it. A word list tells you what a language contains. A story carries its vocabulary too, but also its pronunciation, memory, humour, and grief. A lexicon survives only if there is an ecosystem for it to live in: people who still speak it, day to day. The Nagamese language shows what happens

without that ecosystem. It draws its vocabulary from Assamese and Naga languages, and functions as a lingua franca, market language and a creole that helps trade happen.

Community-centred documentation looks different, as PARI's own story collection shows: Sunita and the Kolami language, Kamlesh, and Bhiklya Dada's first-person account are three examples worth returning to.

Metadata comes in two layers, and a complete record needs both. Conventional metadata holds the facts: the speaker, the language name, the village, the location, and the recording details. Contextual metadata holds the harder things: why the story was told, when and how the community uses it, what it means culturally, and what the community's core memories actually are. Take a PARI story about a farming crisis. Conventional metadata includes interview details, but questions about land, rights, debt, and migration are also part of the language's context because livelihood and language are connected.

The table below sets out what each layer tells you and what you can capture for it:

Layer	What does it tell us?	What can we capture?
Conventional metadata	Where did it come from?	Speaker, language, village, location, date, device, GPS, etc.
Linguistic layer	What is in the language?	Word lists, transcription, translation, IPA-pronunciation, grammar, picture books, etc
Contextual metadata	Why does it matter in people's lives?	Purpose, community use, when and how it is used, memory, relationships, livelihood, cultural/social context, etc.

File naming, arrangement, and searchability form one necessary layer of metadata. Contextual metadata goes further. Done well,

it can reveal a connection between something as dry as census data and something as human as language endangerment.

Field notes:

Ritu Sharma is a linguist, journalist, and wildlife rehabilitator based in Maharashtra. She is a Senior Assistant Editor at PARI and anchors its Endangered Languages Project, which documents India's vulnerable languages through the voices and lived experiences of people who speak them.

In her own family photograph above are seen her grandmother, Nandi Devi Kukreti, with the family's companion dog, taken in 2015. Its technical metadata is straightforward: who, what, when, where. What it cannot tell you is why her grandmother, a Kukreti, migrated to Mumbai with her grandfather when Ritu's father was four years old, around the time of the 1962 India-China war, or why the family afterwards carried the surname Sharma instead. The photo and its technical metadata are accurate and complete, but still cannot answer either question.

A second family detail makes the same point land harder: the dog's own name. It was informally called Pakhu, a Garhwali word that can mean a loyal guard, someone innocent, or someone of the hills. But when Ritu and her brother had to register it on paper, they chose Brownie instead, because Pakhu embarrassed them. Only on a later registration, in 2019, did they write Pakhu. Neither name choice is a fact so much as a small, real record of how a family related to its own language at two different points in time. That is exactly what contextual metadata is for.

Sunita is a cotton farmer in Yavatmal district, Maharashtra, and a fluent speaker of Kolami, a Dravidian language with fewer than 10,000 remaining speakers. Ritu had been told in advance that linguists considered Kolami endangered, and asked Sunita

about it while helping her pick cotton in the field. Sunita looked at her in confusion and said she did not think her language was dying, since she and her community still spoke it. Ritu did not set the two positions against each other. Instead, she noted what any technical metadata would miss: Yavatmal is a district where farmer suicides are frequent, more than 150 in 2026 alone, and Sunita speaks Marathi outside her home because that is the region's market language, the one that puts food on the table, not because she has abandoned Kolami. Ritu went on to film Sunita at work, recording the specific Kolami vocabulary of cotton farming rather than an abstract word list. Ritu's conclusion: contextual metadata does not replace technical metadata, but without it, a recording can look complete while leaving out the exact reasons behind a person's own choices about their language.

7.5 Field example: documenting Tharu healers, tattoos, and the Shama Chakeva festival

Sanjib Chaudhary

This section ties together the ideas above through one documenter's own path, from writing about a community to sustained, community-involved documentation over many years.

The Tharu are Nepal's fourth-largest ethnic group, living in its southern districts along the border with India. They speak several Tharu language variants, including Kochila Tharu (also called Morangiya, Saptariya Tharu, or Mid-Eastern Tharu), spoken by about 250,000 people across eastern and central Nepal.

Sanjib Chaudhary has documented the Tharu community since 2007, when he began blogging about Tharu culture. He started editing Wikipedia in 2016 and contributing to Wikimedia Commons in 2018, documenting traditional healers' cures, Tharu tattoo motifs, and, most recently, the Shama Chakeva festival,

with support from the Wikimedia Foundation.

Shama Chakeva marks the relationship between brothers and sisters. Tharus and Maithils celebrate it together across Nepal's southern plains and the neighbouring districts of India, with clay statuettes, nightly songs, and a closing ritual at a pond or a water source.

His method offers a model worth reusing. He found community resource persons first, the women who already sang the festival's songs, rather than starting from a blank page. He took consent as a community member himself, which changed how people responded to the request. And because songs and stories can vary by region, he recorded the source region and language variant alongside the material itself.

The outcome went beyond a completed archive. The material became a storybook for schoolgoing children. A partnership with an artist taught the women's group to make papier-mache statuettes for sale, sustaining both the tradition and the income of the community members who kept it alive.

Field notes:

Sanjib's mother tongue is Eastern Tharu (also called Saptariya or Kochila Tharu) and is spoken by roughly 250,000 people. He has blogged about Tharu culture since 2007 and started contributing to Wikipedia in 2016 and to Wikimedia Commons in 2018, documenting traditional Tharu healing practices with Subhashish Panigrahi, Tharu women's vanishing tattoo motifs, and, with a Wikimedia Foundation grant, the songs sung during the Shama Chakeva festival.

His own documentation connects directly to Khator's example of Neela Madhaba and Jagannath: in his community, a deity now worshipped as Shambhunath was, in the Tharu telling, Semnath, a person from his own community before being

Sanskritised. Because he was already a community member and spoke the language, finding people willing to share songs and their meanings was straightforward, and consent came easily once he explained that the objective was to document the material for future generations. Working with an artist from the same community to illustrate the story, the first draft went back to the community for review, and elders corrected specific details Sanjib had missed: the central character should not be shown with blue or white skin, but dark skin, and dressed in the clothes the community has actually worn for generations, not a generic depiction.

On sustaining the practice rather than only recording it, Sanjib and the community discussed the clay statuettes at the centre of the festival, worshipped daily until they are immersed in water and discarded. With the women's consent, he worked with an artist to develop a papier-mache version, sturdier and lighter than clay, that could be sold as a souvenir, giving the makers an income while carrying the story to a wider audience. The reach of that decision surprised him. After he posted photos on Instagram, a vendor in Taiwan contacted him to connect him with the women directly, and he is now working with art shops in Kathmandu to take Shama Chakeva's story further.

7.6 Teaching this module

If you only document and archive a language, you can skip this section: it is for teaching others.

Session goal: help the people you're teaching shift from viewing an archive as a one-time act of preserving the past to seeing it as an ongoing practice that serves their community in the present. By the end, each person should be able to name at least one thing beyond a word list worth documenting, and explain why context matters as much as the recording itself.

A few concrete ways to run it:

For §7.1, open by asking the group, or the person you're mentoring, to describe what they imagine of an archive, before you introduce the difference between preserving a language and sustaining one. The Salihan story in the field notes works well as a discussion opener on cultural appropriation, ahead of introducing PARI's own initiative.

For §7.4, open with a photograph you own, the way Ritu Sharma's example does. Build out its technical metadata together, then ask what it cannot tell you. This is a concrete route into the difference between technical and contextual metadata, and it works well one-on-one and in a group.

Across the module, use the field notes as your material. They are written to be told as stories.

The questions below were raised by community members in earlier workshops. In a group, they make a good closing discussion. One-on-one, use them as conversation starters rather than a checklist. None of them has one tidy answer: reason through them using PARI's stories above, rather than in the abstract.

How do we capture cultural gestures (kinesic translation), and how do we annotate this kind of data?

How do we record elderly people respectfully, especially when we cannot ask them to speak louder or more clearly without it feeling rude?

How do we handle storage: what quality should we record at, versus what quality can the community actually access? For example, recording in WAV, but the community can only access MP3.

How do we handle sensitive topics when someone shares a personal story: when to stop, when not to stop, and how to respond?

What should I document about my language? Proverbs, idioms, debates, public speeches?

How do we document a language or archive with marginalised communities, women, and dominant communities together as one group: what should we keep in mind?

If someone gives false, incorrect, or unverified testimony, how do we verify and address it? This also raises a deeper question: whose truth are we talking about?

Notes

Notes

Notes

8. Raising funds for your language documentation & archiving project

Mandana Seyfeddinipur

This module is to help you understand how to fund language documentation work and how institutional archiving works, as two practical skills every community documenter needs, not as a description of one organisation. By the end, you should be able to name at least one route to funding your documentation work, describe in outline what depositing with an institutional archive involves, and explain in your own words why "one archive, one platform" cannot serve every community the same way.

8.1 Fundraising in recording and keeping languages alive

Documentation and archiving cost money for years. You need money for equipment, travel, and data storage. You also need money to pay speakers, translators, and documenters fairly (see Module 2 and Appendix B).

Fundraising is a skill to learn, just like recording and archiving. It is needed for everyone — those working alone, those working with a small group or as an organisation. Community documenters can learn it too.

8.2 Where funding can come from

There are different ways to fund your documentation and archiving work.

Some funders give grants only for documentation. The Endangered Languages Documentation Programme (ELDP)¹ is one well-known example, but it is not the only one. The Firebird Foundation funds oral history recording, and the other funders include the Endangered Languages Fund, the Foundation for Endangered Languages, and the Gesellschaft für bedrohte Sprachen (Society for Endangered Languages), based in

Germany. Each funder gives a different amount of money and has a different focus — some support language revival, others support documentation only. ELDP funds language documentation, not revival. §8.7 lists all of these funders together, with links.

Some funders give training and support along with money, or instead of money — §8.3 explains this using ELDP as an example.

You do not always need a funder. Communities can also raise money locally or get support that is not money. For example, a free space to work in, devices, food, transport, or people's volunteer time (see Module 2.3).

8.3 Applying for a grant

Most documentation grants ask for similar things: what you plan to document, who you will work with, a rough budget, and a plan for where your recordings will be kept later.

ELDP gives two kinds of grants. The Legacy Material Grant is for people who already have old recordings — cassettes, reel-to-reel tapes, mini-discs — made before digital recording was common. ELDP pays to make these recordings digital and to write down their metadata: who is speaking, what they are talking about, and when and where it was recorded.

Field notes:

Mandana Seyfeddinipur is the director of the Endangered Languages Archive (ELAR) and its sister concern, the Endangered Languages Documentation Programme (ELDP), both based at the Berlin-Brandenburg Academy of Sciences and Humanities in Germany.

A 95-year-old Jesuit priest once gave ELAR a set of mini-discs. Decades earlier, he had recorded the creation stories of the

Yekuana community in Venezuela, and he wanted the recordings to go back to the community before he died. It took ELAR's team four weeks just to find a machine that could still play the discs. Mandana's colleague at ELAR found an old player and digitised the recordings; another colleague travelled to the priest's home near Berlin to sit with him and record the metadata before the recordings could be properly archived.

The Documentation Grant is for making new recordings. It covers equipment, such as cameras, audio recorders, and laptops, and the documenter's fees. If you are documenting, you will be spending time and might not be able to do another paid job. So your time needs to be paid for. You can apply each year from June. The deadline is 1 October, and the result will be out in April. If you get the grant, you also get training and ongoing support, usually once a week or once every two weeks, on how to use the equipment and do the work. ELDP works directly with each community member, not with the entire community as a single applicant. Where someone cannot easily buy equipment themselves – for example, some scholars in sub-Saharan Africa without credit cards had to buy expensive equipment – ELDP buys it and sends it to them directly.

One condition applies to every ELDP grant: the recordings must eventually be shared with public access. However, recordings that are culturally sensitive – for example, a women's language only meant to be heard by women – or that could put a speaker at risk, should stay restricted. If a recording could put anyone in danger, do not record it at all. The rule exists mainly to stop researchers from sitting on material until they finish publishing their own work.

The above apply to most grants, including ELDP.

8.4 Institutional deposit — what it involves, and its real limits

See Module 7 for the practical steps of contacting any archive and depositing your files. This section explains something different: what an archive actually does for you, and what it cannot do.

A laptop, an external hard drive, a Dropbox folder, or a video platform like YouTube can hold your recordings, but none of them is safe for the long term. A computer breaks. A hard drive gets lost or damaged. A cloud password gets forgotten. A commercial platform can simply delete your content if it decides to. Each time, a recording that may be the only record of a community's knowledge is lost forever. A professional archive's job is to protect against exactly these risks. It also keeps moving your recordings into newer file formats as old ones stop working, so something recorded today can still be played after sixty years.

ELAR began in 2002 and now holds recordings in over 700 languages, free to access. A community can choose one of three access levels for their own recordings: Open Access, where anyone can use the material for anything, even commercially; Public Access, where only registered users can see it, and only for education, not commercial use; and Closed Access, where anyone who wants to see a recording must write to ELAR and ELAR then asks the community's own steward for permission first. Other archives include CoRSAL, an archive for Indian languages, and several language archives that are a part of the DELAMAN network. They all offer this same kind of graded access to protect communities' privacy and wishes.

Field notes:

People sometimes find it hard to use institutional archives for

two reasons: they can be hard to use if you don't already know how archive websites work, and they can be built around ways of organising knowledge that are not the community's own. For example, a typical ELAR deposit page is laid out in "global north ways": left to right, top to bottom, and with metadata written for linguists, not for the community whose recordings it holds. It's not easy to manoeuvre, it's not easy to find things in there. But the archive cannot do everything. It can promise that your recordings are safe and will still be there decades from now. But browsing it might not feel easy or familiar (see §8.5). An archive must also protect its materials from misuse. ELAR has faced repeated attempts by bots to copy, or "scrape," its collection. ELAR recommends that any community worried about this move their materials from Open Access to Public Access, so anyone wanting to use the recordings to train an AI system has to contact the depositor directly and ask first.

8.5 No single institution or platform can fit every community

No single platform, archive, or funder can serve every community's needs. Any digital system, whether a language archive or a project like Wikipedia, carries its own way of organising knowledge. None of them is neutral.

Different communities want different things. Some want a YouTube channel with subtitled videos. Others do not care about video recordings at all and want a printed dictionary or a book instead.

An archive is an archive, not a community platform. It can help you archiving, but it cannot build the bridge to your own community — a YouTube channel, a Wikipedia page, a printed book. That work belongs to you, the documenter, and your community. We cannot do what you do at all.

Field notes:

Sharing is not the same as preserving. A recording on social media and a recording in an archive solve two different problems. The ELAR team now trains communities to document using ordinary phones. An Android phone is already logged in to a Google account, so a community member can easily upload videos to a private channel shared only with their community. If a community wants, an archive can then copy that same video for long-term preservation. That way, the recording is both shared in a way the community actually uses (social media) and archived in a way built to last.

ELAR ran this exercise in Vanuatu at a summer school. About 50 Ni-Vanuatu participants, from linguistics students to village chiefs, worried that young people no longer fished, taking with them the knowledge of fishing and the sea's stories.

Participants used a phone dictionary app to photograph and name objects, trees, and plants. It took them a few seconds, and they wouldn't come back because they enjoyed it so much.

They eventually built a working dictionary within the workshop itself, with the visiting linguists sitting at the back to help only if asked. The group then recorded a video together, uploaded it with basic details, and the ELAR staff copied the files into the archive overnight. By morning, the community had both a YouTube channel and an archived collection.

A second example: an Iraqi Jewish scholar who had left Iraq decades earlier made recordings he wanted to share with others, his own age, most of them in their 80s and 90s. Many of them could not manage ELAR's sign-up process without a grandchild's help. So he archived everything with ELAR for safekeeping, then made a YouTube channel and sent that link to the people he actually wanted to reach. As Mandana says, "Now he has the best of both worlds."

One problem is not yet solved: an archive cannot check if a person requesting access to a file truly belongs to the speaker community. Amina Mettouchi, a linguist who works with Amazigh (Berber) communities, once asked Mandana where community-led recordings could be archived, on the condition that only community members could see them. Mandana answered, “We cannot check a Tamazigh passport.” An archive has no way to confirm who truly belongs to a community. No institution has fully answered yet who should be allowed to see a deposit, and how a community’s consent is worked out in the first place.

8.6 Sharing copies with the community, and planning follow-up support

Module 7.4 explains the practical steps of sharing copies with a community. This section covers what makes that work in practice: consent and simple metadata.

Consent, not technology, is the hardest part of sharing and archiving a community's recordings. Keeping a recording in an archive and sharing it publicly both need real permission first, and documenters sometimes assume this is fine to do without asking. We don't know how many people have consented to have their video recordings on YouTube.

An archive only needs four simple details to make a recording findable: where it was recorded, when it was recorded, who is speaking, and what it is about. You do not have to collect this separately — a documenter can simply say it out loud at the end of a recording: today's date, the place, who they spoke to, and what the conversation was about. Someone can write this down later from the recording itself.

Field notes:

The COVID-19 pandemic changed language documentation work. When ELAR's own staff could not travel, a community in Vanuatu began uploading their own recordings to YouTube and writing basic captions for them — work ELAR's linguists had done before. Mandana's colleague, Vera, working 12 hours away, would log in each night, download what the community had uploaded that day, and add it to the archive. By the next day, both a YouTube copy and an archived copy existed.

8.7 Teaching this module

If you only document and archive a language, you can skip this section: it is for teaching others.

Session goal: Give a group a real, concrete sense of what an institutional deposit looks like from the inside, and a chance to ask their own questions about funding and archiving before they have to figure it out on their own.

A few concrete ways to run it:

Share your screen and walk through an actual deposit or catalogue entry live, using an archive you know as the example. Encourage the participants to ask questions after. People respond better to seeing a real entry than to a description of one.

If they are quiet at first, three starter questions tend to get things going: what happens to materials after they are deposited, who can access them and whether that can change later if the community changes its mind, and what the very first email or message to an archive should say.

8.8 Resources

Funders and archives named in this module:

Endangered Languages Documentation Programme (ELDP) grants for language documentation fieldwork worldwide, plus Legacy Material Grants for digitising old recordings. See <https://www.eldp.net/>.

Endangered Languages Archive (ELAR) is a free, long-term archive for language documentation collections. See <https://www.elararchive.org/>.

Firebird Foundation for Anthropological Research grants up to \$10,000, including a supplemental grant for oral literature and traditional ecological knowledge; deadlines fall on 15 January and 15 June each year. See <https://www.firebirdresearchgrants.org/>.

Endangered Language Fund (Language Legacies grant) is open to community members and researchers from any country. See <https://www.endangeredlanguagefund.org/>.

Foundation for Endangered Languages gives small grants, around \$1,000, for documentation and revitalisation projects. <https://ogmios.org/>.

Notes

Notes

A Note on Language

Words carry judgement, and what someone intends does not matter. This manual chooses its words on purpose, and this note explains why, so the choice does not sit unexplained on the page.

Adivasi, tribal, local language, mother tongue. This manual uses these words for the communities and languages it documents, not "vernacular," "minority," "low-resource," or "dialect" ("boli,"). Those words measure a language against another one treated as the real, default language, usually a state or colonial language. A mother tongue is not a lesser version of anything. It is the language a community actually thinks, sings, and prays in.

Dalit, Savarna, Brahminical, Dalit-Bahujan. This manual uses "Dalit" for the communities designated under the Indian constitutional term Scheduled Castes (SC), who were historically deemed "untouchable" under the caste system, and "Savarna" for dominant-caste communities, instead of "upper-caste/lower-caste" or "forward/backward caste." Those framings restate a hierarchy as though it were a neutral fact. "Brahminical," used elsewhere in this manual (§7.1) for the caste power that turned the deity Neela Madhaba into Lord Jagannath, names that same system. Caste slurs, "chamar," "chamaar," and "Harijan" among them, appear nowhere in this manual except inside a direct quote, in someone's own words. "OBC" (other backward castes), a constitutional category, is kept as it is.

Women, men. This manual uses "women" and "men," not "female" and "male." "Female" and "male" describe biological sex, which is not discussed in this manual. This manual discusses gendered work, gendered risk, and gendered access to safety and technology, better described in the words communities use for who they are.

Accessibility, disabled. This manual recommends using

"accessibility" and "disabled," including in Appendix A's guidance on captioning for deaf and hard-of-hearing viewers, not "divyang," the term used by the Indian government, or "specially abled," used in other contexts. The disabled people themselves reject such terms for covering up real conditions and not addressing issues.

Word choices in this manual work the same way: they are a deliberate stance. A community's own preferred word, where it differs from anything above, always takes precedence over this note.

Appendix A: Text Style Guide

This guide is for writing down spoken words in language documentation: as captions on screen, as subtitles in another language, or as a full written transcript. Captioning writes what is said, and important non-speech sounds, as on-screen text in the language actually spoken, mainly for deaf and hard-of-hearing viewers. Subtitling translates those captions into a different language. Transcribing writes the spoken words down as a plain document, without time codes.

A.1 Captioning

Captioning helps people who are deaf or hard of hearing read what they cannot hear. It only works for languages that have a script, and for viewers who can read that script. If a language has no script of its own, you can still caption it using the script of a neighbouring language taught locally in school.

Accuracy. Write what you hear: captions should match the spoken words exactly, including slang and informal language, if that is how the speaker actually talks. If a speaker uses words from a different language, set those words in *italics*.

Timing and format. Keep each line to around 30–35 characters so it stays easy to read; break a long sentence across two lines, or into two or three meaningful parts, rather than cramming it onto one line. "We are keeping the crops in a closed bin in winter," for example, reads better split as "We are keeping the crops" and "in a closed bin in winter." Never show more than two lines of text on screen at once. Time each caption to appear when the speaker begins and disappear when they stop, and keep it on screen long enough to read comfortably, between three and seven seconds. Where a sentence is long, break it where the speaker naturally pauses, never in the middle of a phrase.

Giving background. Use brackets for sounds that matter to the

story, for example, (DRUMMING) or (LAUGHING). When more than one person is speaking, name the speaker at the start of their line; if a full name is long, shorten it to a first name plus a surname initial, "Surendra S.P." rather than "Surendra Singh Pangtey," and note the full name separately in the recording's description text.

A.2 Subtitling

This subtitle convention is based on the BBC Subtitle Guidelines¹ (Version 1.2.5, March 2026), itself based on the EBU-TT-D standard.² It covers multilingual interviews, where an interviewer and interviewee speak different languages or switch between languages within the same recording.

File format. Use the SRT (.srt) or VTT format for subtitles. Wikimedia Commons uses TimedText,³ which is close enough to SRT that you can open your SRT file in a plain text editor and copy its contents straight into TimedText. Each subtitle block needs a sequential index number, a timecode in the format HH:MM:SS,mmm -> HH:MM:SS,mmm, one or two lines of subtitle text, and a blank line separating it from the next block:

```
1
00:00:01,292 -> 00:00:02,001
```

Your name?

```
2
00:00:02,210 -> 00:00:03,086
```

Sukra Dhangdamajhi

Language layers. Subtitle each recording in at least two layers: the source language, the primary language spoken, transcribed as closely as possible, including code-switches (see "Code-

switching" below); and a translation language, a more widely spoken language the source subtitles are translated into. Keep each layer as its own SRT file, named LanguageCode-RecordingID-LayerLanguageCode.srt — for example, Bfw-Munaremo-SukraDhangdamajhi.or.srt for the Odia (or) layer of a Bonda (Bfw) recording. Use the ISO 639 code for languages that have one; for languages without an assigned code, use the Glottolog⁴ code or the language's full name. Dialects are not usually captured separately in subtitles.

Identifying speakers. The first time each speaker appears in a subtitle block, name them in full, on a separate line above their speech, followed by a colon and a line break. After that, only repeat their name when the speaker changes, or after a gap of 30 seconds or more:

1

00:00:01,292 -> 00:00:02,001

GOBARDHAN PANDA:

Your name?

2

00:00:02,210 -> 00:00:03,086

SUKRA DHANGDAMAJHI:

Sukra Dhangdamajhi

Off-screen speakers. When you can hear a speaker but they are not visible on screen, put a single quote before and after their speech, for example, 'What all do you do?'. If the off-screen speech runs across several lines, keep one quote at the very start and one at the very end.

Inaudible speech. When speech cannot be heard, explain why, in

capital letters if you are writing in Latin script: POURS WATER INTO GLASS. INAUDIBLE SPEECH.

Whispering. Label whispered speech the first time it occurs, WHISPERS: I knew it. and drop the label afterwards if the whispering continues. For a long whispered subtitle, put brackets around the whole line instead: (We often pickled raw mangoes in summer and ate throughout the year.)

Code-switching. Code-switching is when a speaker shifts from one language to another within the same speech, whether mid-sentence or across part of a recording; it is common wherever interviewers and interviewees share a language other than the one being documented. Name the language in capital letters where the switch happens:

IN NEPALI: Our worship rituals are known to the elders.

For a switch mid-line, mark it inline in angle brackets at the point it happens:

We speak Raji, {NEPALI}, during our worship rituals

Use a line break once the speaker returns to the primary language. Name the speaker, the first time they appear, alongside the language tag if an off-screen speaker code-switches:

UDAY AALEY: And in which places {NEPALI} do you use the Raji language?

Only mark a genuine language switch this way, not a single loanword, a place or a person's name.

Hesitations, fillers, and overlapping speech. Keep meaningful hesitations, such as "Hmm..." or "Okay...", but do not transcribe every filler if one recurs often. Too many on-screen makes a subtitle unreadable.

Indigenous and community-specific terms. Where a term has no real translation into another language, keep the original word, and set it in capital letters to mark it apart from the rest of the subtitle: I am the DISARI of this village.

Line length and reading speed. Recommended:

Maximum two lines per subtitle.

Maximum 42 characters per line, including spaces.

Minimum display duration: one second.

Target reading speed: 160–180 words per minute for a general audience; allow more time for recordings with technical, cultural, or Indigenous vocabulary.

Break lines at a natural pause, especially where a sentence must split mid-way – for example, "We built this small house" and "to keep our domestic animals."

Metadata for each subtitle file. Recording this alongside each SRT file makes the collection usable later. For Wikimedia Commons, add it to the file's own information page. Keep:

Primary language: the language spoken, with its ISO code.

Subtitle language: the language of this particular SRT file.

Speaker(s) and interviewer(s): full names, where possible.

Subtitler(s): full names and method, for example, "transcribed from Bonda by X; translated to Odia by Y; English translation and Odia editing by Z."

Code-switch language(s), if any, with their codes.

Notes: anything else worth recording, including any departure from this convention.

5 Wikimedia Commons, OpenSpeaks file information template, <https://commons.wikimedia.org/wiki/Template:OpenSpeaks>

1. BBC Subtitle Guidelines, Version 1.2.5 (March 2026), <https://www.bbc.co.uk/accessibility/forproducts/guides/subtitles>
2. EBU-TT-D Subtitling Distribution Format (Tech3380), <https://tech.ebu.ch/publications/tech3380>
3. Wikimedia Commons, Timed Text, https://commons.wikimedia.org/wiki/Commons:Timed_Text
4. Glottolog, <https://www.glottolog.org/>

Appendix B: Oral Knowledge Framework

This framework is for anyone archiving oral knowledge: how to plan a recording, how to work with a community to discuss and agree on what happens to it, and how to document that process itself. It builds on OpenSpeaks Archives' earlier work, a 2024–2025 pilot, a 2025–2026 implementation phase, and case study analysis.¹

B.1 FAIR and CARE principles

This framework speaks to three kinds of documenters: a community member recording their own language and culture, a neighbouring Indigenous documenter working with a language close to their own, and an external non-Indigenous documenter, such as a researcher, volunteer, NGO worker, or Wikimedian. The steps below are broadly the same for all three, but responsibility and power are not distributed the same way, so each role carries its own extra considerations, covered further down.

Two sets of principles run through this framework. FAIR, Findable, Accessible, Interoperable, Reusable,² helps organise recordings and metadata, so they stay useful now and later. CARE, Collective Benefit, Authority to Control, Responsibility, Ethics,³ makes sure the recording process and its output respect Indigenous rights, community priorities, and lived experience.

FAIR is about making digital files and their descriptions easy to find, access, combine, and reuse. Findable means people can discover a recording through a simple search, because it has a clear title, language name, and keywords. Accessible means that once someone finds a file, they know how to request or access it, whether access is open or restricted. Interoperable means files and their metadata use formats and vocabularies that different tools and platforms, including Wikimedia projects, can actually work with. Reusable means a file carries a clear licence and enough description of its own origin that someone else can reuse

it without guessing.

CARE is about people and relationships, centring Indigenous and other marginalised communities alongside the data itself.

Collective Benefit means documentation should help the whole community first, and the project, researcher, or platform behind it second. Authority to Control means the community decides what is recorded, how it is described, how it is shared, and who can reuse it. Responsibility means a documenter acts with care, follows local protocol, and supports the community's own capacity to do this work. Ethics means documentation and its reuse respect dignity, safety, and local values; not everything that can be published should be published.

Before you record: planning with FAIR and CARE

These five steps apply to whatever your role:

- 1. Talk first, record later.** Explain who you are, why you want to record, and how the recordings may be used. Ask what the community itself wants from the documentation, teaching materials, an archive for younger people, or citations for Wikipedia, among other things.
- 2. Agree on topics together.** Prioritise what the community actually cares about: traditional knowledge, biodiversity, stories, songs, local history, or everyday conversation. Avoid sensitive or private topics unless you have clear, informed consent and local approval.
- 3. Decide together what can be public and what cannot.** Some content, sacred knowledge, restricted rituals, or anything that could cause harm, may never be meant to go online at all. Other content can be shared only under conditions: within the community only, only after a certain time, or only without naming real individuals.

4. Choose consent and release forms. Use simple, translated consent forms that explain who is recording, what will happen with the recordings, how the interviewee can say no or withdraw later, and which platforms may host the files, for example, Wikimedia Commons or a local archive.

5. Plan how the community will benefit. Decide in advance how you will share back: copies of the recordings, subtitles, local screenings, training, or a community-controlled archive. Where there is funding, agree on fair remuneration, and keep part of the project budget for the community's own priorities.

Extra responsibilities by role

A **community member** is documenting from within. You likely already know local protocol, but it is still worth asking elders and others what is appropriate for this particular recording. Plan how other community members will be involved in reviewing, subtitling, and using the material, so the benefit is shared, not kept by one person.

A **neighbouring Indigenous documenter** may share a similar history of oppression and erasure with the community they are working with, but is still a guest there. Make sure local leaders and elders invite you and guide the process, and pay attention to how protocol differs: do not assume your own community's practices automatically transfer.

An **external non-Indigenous documenter** carries more responsibility because of the power difference between you and the community. Do not design the project alone: co-design it with community members, share decision-making with them, and be ready to step back if the community does not want a particular kind of documentation or publication.

B.2 Practical steps during recording and after

Ethical recording practices (CARE)

Explain again before you record. Repeat the key points, the topic, the purpose, where the recording may be shared, and the interviewee's right to stop at any time.

Respect pauses and refusals. If someone does not want to answer or seems uncomfortable, change the question or stop altogether. Not answering is always an acceptable answer.

Co-create the questions. Where possible, ask the interviewee or community members what they want to be asked, and let them suggest topics, corrections, or a different angle.

Include context. Record a short segment on who is speaking, where, when, and why the topic matters. This helps whoever interprets or reuses the recording later.

Technical practices (FAIR)

Use clear, easy-to-understand file names. For example: LanguageCode_Village_Topic_SpeakerInitials_YYYYMMDD.extension.

Record good audio and video. Use a stable device, minimise background noise where you can, and make sure everyone who wants to be clearly seen or heard actually is, unless they have asked not to be.

Capture basic metadata while recording. Note the language or languages, the location (at whatever level of detail the community is comfortable sharing), the date, and the key themes, whether in a notebook, a form, or an app.

Think about accessibility from the start. Where you can, record in a way that leaves room to add subtitles, transcripts, or translations later; shorter segments are usually easier to subtitle and reuse.

After recording: describing, reviewing, and deciding access

Community review (CARE):

1. Community-first review. Before uploading anywhere, share the recordings with the interviewees and, where appropriate, community reviewers. Ask directly: is this accurate, is anything missing, is anything you want removed or kept private?

2. Record the decisions. Note which recordings are public, restricted to the community, delayed until a later time or event, or never to be shared at all.

3. Discuss how to handle sensitive material. Agree on how to treat names, faces, sacred sites, or personal stories, whether that means blurring faces, changing names, or keeping only the audio.

Metadata and licences (FAIR):

Write rich but respectful descriptions, including language, a content summary, keywords, and any cultural context the community has approved. **Use licences the community has actually agreed to:** for Wikimedia, that often means Creative Commons Attribution-ShareAlike (CC BY-SA), but only once the community understands and accepts what that implies; a community may allow some recordings under CC BY-SA and not others, and that choice should be respected file by file. **Connect recordings to other data** where it makes sense, such as related recordings, Wikipedia articles, or other references, since this makes a collection more interoperable and reusable. **Keep**

metadata even when a file itself stays restricted: a minimal description, with the community's approval, lets others know the resource exists and how to request access, without exposing the file itself.

Using recordings in Wikimedia and other platforms

As a **community member**, decide together which recordings can go to Wikimedia Commons or another open repository, and use subtitles, transcripts, and translations to make the content usable across different Wikimedia projects: Wikipedia articles, Wikidata items, or Wikisource, where a transcript has been peer-reviewed. Treat Wikimedia as only one layer of your archive; keep or build a community-controlled copy first.

A **neighbouring Indigenous documenter** should check again that publishing on a global platform still matches what the recorded community actually wants. Where you are acting as a bridge to Wikimedia, keep local leaders involved in decisions about article content, image choice, and framing, and share back which pages or entries were created or improved because of the recordings.

If you are an **external non-Indigenous documenter**, make sure local learning resources, language classes, or community archives are part of the project design from the very beginning. Do not rush to put everything online: prioritise what the community explicitly wants on Wikimedia, and what they feel is framed correctly. In academic or public writing, cite the community and interviewees properly, and share drafts with them for feedback before publishing, wherever that is possible.

Role-specific checklists

If you are a community member:

Ask elders or leaders what to document and what to avoid.

Use simple consent explanations in your own language.

Plan how the recordings will be used to teach younger speakers.

Keep copies in safe, community-controlled storage.

Decide together which files go to Wikimedia and which stay internal.

If you are a neighbouring Indigenous documenter:

Ask to be invited; do not assume you should document.

Learn and follow local protocol, even where it differs from your own.

Share equipment and skills, alongside files.

Make sure decisions about access and licensing rest with the documented community.

Be ready to hand over the archive if the community prefers to manage it directly.

If you are an external non-Indigenous documenter:

Co-design the project with community members from the very start.

Budget for fair compensation, training, and community-controlled infrastructure.

Use consent and governance tools that are readable and translated.

Plan for your own exit: how will the work continue without you?

Be accountable: report back, share findings, and stay open to criticism.

B.3 Fair, equitable remuneration

This compensation model rests on the FAIR and CARE principles above, together with three further ideas. The Equity-Centered Participatory Compensation Model⁴ centres marginalised voices and fair pay. Reparative justice means acknowledging historical oppression and the systemic barriers it left behind. Intersectionality means recognising that a person can face more than one kind of marginalisation at once, and that these compound rather than simply add up.

The table below gives example rates only. Fix the actual amount to your own local labour system and cost of living; do not copy these numbers directly.

Role	Base rate (USD/hour)	Reason
Native speaker interviewee	\$25–35	Knowledge holder; primary contributor
Subtitle creator or translator	\$20–30	Skilled linguistic work; bilingual or multilingual
Community reviewer	\$20–25	Quality assurance; cultural verification
Language coordinator	\$25–35	Project management; community liaison
Basic technical support	\$18–25	Basic technical help

Intersectional adjustment factors

A person can carry more than one layer of marginalisation at once. Treat these as compounding, not as separate line items to add up by hand: someone who is both Dalit and a woman, for instance, faces both barriers together, and a fair rate should reflect that combination, not just the larger of the two alone.

Three layers matter most. Historical marginalisation covers caste, tribe, race, and religion, for example, a Dalit, Adivasi, or OBC contributor, or someone from a persecuted religious minority such as Dalit-Christians in India, Muslims in India, Myanmar, or the United States, Christians in Pakistan, or Tamil Hindus in Sri Lanka. Gender-based marginalisation covers women and femme-identifying contributors, trans and non-binary contributors, and widows working in patriarchal contexts. Disability and health covers physical and sensory disability, neurodivergence, and chronic illness. Each layer typically raises the base rate by between 15% and 30%, weighted by how severe and long-standing the particular barrier is.

Use the base rate table as your starting point, then raise it deliberately for each layer a person carries, using the categories above as a guide to which factors matter and roughly how much weight each deserves. Work out the exact figure with a dedicated calculation tool built on this same approach, rather than by hand from a printed table. What this appendix gives you is the principle: pay more for compounded marginalisation, never less, and never flatten someone's circumstances into a single rate that ignores what they are carrying.

B.4 Community assessment of oral histories

A community reviewing its own recordings of an Indigenous language can work much like peer review does in scientific publishing: a protocol for checking, correcting, and approving what has been recorded. Documenting this process is worth doing for its own sake. Record which community members reviewed the material, what concerns or corrections they raised, how any disagreements were discussed and resolved, and what the group finally agreed on. Kept alongside the recording, this becomes evidence that the material was not only made but also

checked and approved, in the same way editorial review notes accompany a conventional publication.

The assessment itself has to follow local protocol and local ways of knowing: who is allowed to review particular material, what must stay private, and how authority and expertise are actually recognised in that community.

B.5 Multilingual transcription

The goal here is to prepare oral knowledge recordings so they can be cited. Where resources allow, give each recording a transcript and subtitles in at least one widely used language, ideally a neighbouring official language that native speakers are already comfortable with; this lets people outside the community understand the content too. See Appendix A for the practical mechanics of captioning and subtitling.

Transcription quality needs review as well: community members can check a transcript and its subtitles the same way they review the original recording (see B.4 above), adding another layer of verification.

For citation, a time-coded transcript can work much like page numbers do in a printed text: each statement can be linked to a specific moment in the recording. This gives editors and readers a concrete way to verify what was said and where, and pushes back against the idea that oral sources are somehow less precise than written ones.

Use this framework throughout your documentation work, from the first conversation with a community through to sharing recordings publicly.

1. Subhashish Panigrahi, Opino Gomango, and Kimmi Pal, "OpenSpeaks Archives: Citing Low-Resourced Language Oral History Multimedia," Wiki Workshop 2026 (Wikimedia

Foundation, 25 March 2026), https://wikiworkshop.org/2026/paper/wikiworkshop_2026_15_openspeaks_archives_citing_lowresourced_language_oral_history_multimedia

2. GO FAIR, "FAIR Principles," <https://www.go-fair.org/fair-principles/>

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

4. Maya Rabinowitz, Isabelle González, and Rhianna C. Rogers, "The Equity-Centered Participatory Compensation Model (EPCM): A Tutorial for This Emergent Methodology," *Journal of Participatory Research Methods* 5, no. 1 (2024), <https://jprm.scholasticahq.com/article/92362-the-equity-centered-participatory-compensation-model-epcm-a-tutorial-for-this-emergent-methodology>

archiving the present

a community guide to
archiving languages,
memories & knowledges

PUBLISHED BY



SUPPORTED BY



Samāgata
Foundation



IN PARTNERSHIP WITH



Endangered
Languages
Project



ENDANGERED
LANGUAGES ARCHIVE



LICENCE



Text and illustrations are licensed CC BY-SA 4.0: copy, adapt, and
reuse them, with credit, under the same licence.



ISBN 9 788198 228901