

University of Montana

ScholarWorks at University of Montana

CoLang 2022 Workshops

CoLang 2022

August 2023

02. Archiving for the Future (A, D, E)

Jaime Pérez González

Susan Smythe Kung

Follow this and additional works at: https://scholarworks.umt.edu/colang2022_workshops

Let us know how access to this document benefits you.

Recommended Citation

Pérez González, Jaime and Kung, Susan Smythe, "02. Archiving for the Future (A, D, E)" (2023). *CoLang 2022 Workshops*. 2.

https://scholarworks.umt.edu/colang2022_workshops/2

This Article is brought to you for free and open access by the CoLang 2022 at ScholarWorks at University of Montana. It has been accepted for inclusion in CoLang 2022 Workshops by an authorized administrator of ScholarWorks at University of Montana. For more information, please contact scholarworks@mso.umt.edu.



COLANG 2022

Archiving for the Future

Susan Smythe Kung

Jaime Pérez González

Land Acknowledgement

The University of Montana acknowledges that we are in the aboriginal territories of the Salish and Kalispel people. Today, we honor the path they have always shown us in caring for this place for the generations to come.



Links

These slides: <https://bit.ly/colang2022-AFTF>

Archiving for the Future online course:

- <https://bit.ly/AFTF-OER> or
- <https://archivingforthefuture.teachable.com/>



Archiving for the Future: Simple Steps for Archiving Language Documentation Collections

An Open Educational Resource (OER) available at
<https://archivingforthefuture.teachable.com/>



Meeting schedule

Day 1

- Introductions to each other
- Introduction to the course
- A very brief history of archiving
- The Simple Steps
- Phase 1: Step 1

Day 2

- Phase 1: Steps 2-4

Day 3

- Phase 2: Steps 5-6

Day 4

- Phase 3: Steps 7-9



Day 1

- Introductions to each other
- Introduction to the course
- A very brief history of language archiving
- The Simple Steps
- Phase 1: Step 1



Facilitator Introductions

Susan Kung, skung@austin.utexas.edu



Jaime Pérez González, locosinguero@gmail.com



Participant Introductions

- Attendance
- Getting to know each other activity



Archiving for the Future (AFTF) - the course, ...

... is NOT Intended to teach participants how

- to archive materials in any specific repository or to explain differences between archives or workflows,
- to digitize analog materials, or
- to build a digital repository.

... is intended to teach some basic data management practices that will

- help you keep your data (**gifts, recordings, curricula**) and metadata (**meta-description**) organized,
- facilitate their ingestion into ANY digital repository or archive, and
- facilitate their current and future discoverability and reuse.



Simple steps for archiving language documentation data

PHASE 1 BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

PHASE 2 DURING

STEP 5
collect
metadata

STEP 6
evaluate
your files

PHASE 3 AFTER

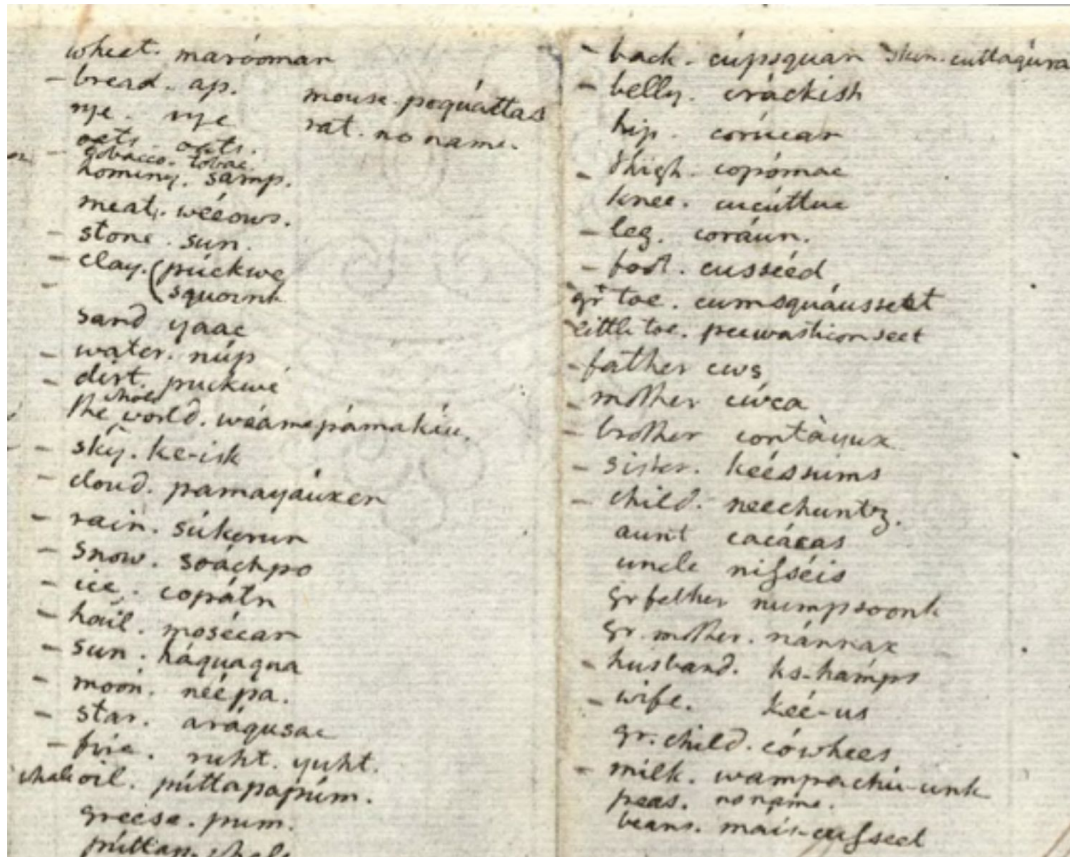
STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

STEP 9
describe your
collection



A brief history of archiving language documentation



Jefferson, Thomas. 1791.
Vocabulary of the Unquachog
Indians. [image] Mss.497.V85.
American Philosophical
Society.

<http://diglib.amphilsoc.org/islandora/object/vocabulary-unquachog-indians>,
Public Domain.



Archiving

Late 20th / early 21st century

- The Digital Divide - the transition from analog to born-digital
- More emphasis placed on the need to ensure that language documentation gets archived (versus being held in private collections and/or thrown out)
- Archiving becomes a key component in language documentation activities
 - Collections created today are multimedia, born-digital, and include large numbers of (sometimes quite large) digital files
 - Creators of materials → curators of their materials
 - Archiving happens for each project and not just at the end of someone's life



Discussion

- What kinds of Indigenous Language Documentation and/or Revitalization (IRD/ILR) materials do you create?
- What vocabulary terms do you use to name those materials?
- How do you feel about the word "data"? [The class discussed this and the word *gift* was offered as an alternative; etymology= *data* is plural of *datum*, which is Latin for 'gift'. Other suggestions include *recordings, curriculum/a*]
- Are the materials (you work with or create) digital? Analog? Both?
- How do you use the materials? Why did you create them? What are their purposes?
- What is the future of those materials?



PHASE 1
BEFORE

STEP 1
choose your
archive

Step 1: Choose your archive



PHASE 1
BEFORE

STEP 1
choose your
archive

Definitions

Primary Data (Gifts): audio or video recordings or written observations of spoken or signed language that are used for analysis or to create other products; e.g., narratives, oral histories, elicitation, experimental protocols. (Different from Himmelmann 2012)

Secondary Data (Gifts): refers to transcriptions, translations, morpheme breakdowns, glosses and other types of annotation that require some level of preliminary analyses of primary data. (Different from Himmelmann 2012)

Metadata (metadescription): information about the data (gifts), including information about what the data (gifts) are, the format of the data (gifts), and their purpose.



PHASE 1
BEFORE

STEP 1
choose your
archive

Purpose of language archives

- Responsibility to the people they serve to make materials available for the long term.
- Long-term *preservation* (making sure the data (gifts) last a long time)
- Long-term *accessibility* (making sure users can access them now and in the future).
- Commitment to migrating the data (gifts) and metadata (metadescription) to new formats.
- Provide citable public versions of stable data (gifts) and datasets (recordings, curricula) that can be used for education, research and other purposes.
- Not meant to be temporary file storage or social media platforms.



PHASE 1
BEFORE

STEP 1
choose your
archive

Repositories vs. Social Media & Storage Platforms

Repositories

- DELAMAN archives (<https://www.delaman.org/>)
- Data repositories
- Physical archives

Storage or Social Media Platforms

- Media Storage Platforms (e.g., Google Drive, Box, DropBox, OneDrive)
- Websites or Streaming Services (e.g., YouTube, Vimeo, SoundCloud)

Major Differences:

- Digital repositories or archives are not intended to be used as temporary storage.
- Archived data must be stable so that the information in the files can be cited.
- Archived data is unlikely to be deleted, and take-down procedures can be difficult, especially when preservation copies are distributed across many locations.



Day 2

Phase 1:

- Finish Step 1
- Steps 2-4



Self-reflection Questions

1. Why are you participating in this workshop?
2. What do you hope to get out of it?
3. What kinds of things are you considering archiving?
4. What do you think might be some *positive outcomes* if you archive the digital multimedia files that you create (either by creating born-digital files or by digitizing analog materials)?
5. What do you think might be some *negative outcomes* if you archive the digital multimedia files that you create?
6. What do you think might happen to the files/materials if you do not archive them?
7. Who do you think will want to use them now or in the future? Can you envision some possible users and their use cases?



PHASE 1
BEFORE

STEP 1
choose your
archive

Repositories vs. Storage Platforms

Repositories (e.g., DELAMAN archives, Data Repositories)

- **Stable** distribution platforms in which files are not likely to change or be deleted & material can be cited.
- There is an institutional commitment to longevity.

Media Storage Platforms (e.g., Google Drive, Box, DropBox, OneDrive)

- Files are highly likely to be moved, deleted, changed or edited.
- Commercial services are subject to frequent changes or discontinuity of service.



PHASE 1
BEFORE

STEP 1
choose your
archive

Repositories vs. Social Media Platforms

Repositories (e.g., DELAMAN archives,
Data Repositories)

- Less easy to access or navigate, but
- Committed to long-term preservation

Websites or Streaming Services (e.g., YouTube,
Vimeo, SoundCloud)

- Easy to access and navigate
- No promise of long-term preservation



PHASE 1
BEFORE

STEP 1
choose your
archive

Picking an Archive

- Are you required to use a specific archive?
- Is there a community-based archive?
- Do you already have a relationship with a specific archive?
- Is there a DELAMAN archive that would be a good fit for the language?
- Is there an areal archive that would be a good fit for the language?
- Will some or all of the language data have to be restricted? What are the repository's policies regarding public vs. restricted data?
- Can the language community access the archive?
 - If not, is there a regional archive or institution that community members can access?



PHASE 1
BEFORE

STEP 1
choose your
archive

Prioritize community access!

- Seek out an archive that the speech community can access and maybe even control access to the collection.
- Establish community access through a Memorandum of Understanding.
- Immediately share copies of the collected materials with people who contributed to their creation.
- Establish a physical or digital archive within the community (e.g., Raspberry Pie, Jukebox)
- Share a digital or physical copy of the materials with an institution that the community members can access, e.g., a school, library, archive, museum, governmental office or cultural program in the community's town, region, state, country
- Set up a local Raspberry Pi for nearby smartphone access (see Thieberger & La Rosa's presentation in ICLDC 7, session 2.1)
- Bear in mind that the easiest way to provide access might be to put video files in YouTube.



PHASE 1
BEFORE

STEP 1
choose your
archive

Questions about Step 1?



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

Step 2: Name Your Files



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

Filenaming Methods

Sequential file naming: Filenames are generated automatically by a recording device and in the sequential order of their creation. Parts of the filename are not meaningful to the context of their creation.

Semantic naming: Parts of the filename are meaningful. Some descriptive information relevant to the context is included in the filename.

Combination naming: Combines semantically meaningful name parts with sequential ordering.



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

Filenaming Tips

1. **Be brief** - Abbreviate filenames and put additional info in the metadata log.
2. **Pay attention to characters** - Use only letters, numbers, dashes, and underscores. Do not use special characters like accent marks or parentheses.
3. **Do not use spaces** - Instead use dashes, underscores or CamelCase.
4. **Do not use periods** - Use a period only to set off the file format from the rest of the filename.
5. **Follow the international archival standard for dates** (YYYY-MM-DD)
6. **Use leading zeros** - Number files in a series sequentially using leading zeros for the single-digit numbers.
7. **Practice versioning** - Distinguish different versions with the letter "v" for version and the version number. Don't forget the leading zero!



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

More Filenaming Tips

- Carefully consider your data and establish a filenaming convention that works for you, and then ...DO NOT CHANGE IT!!!
- Filenames must be consistent over time, and changing a system mid-project will lead to endless issues.
- You will be tempted, but just don't do it!!!



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

Questions



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

Step 3: Pick Enduring File Formats



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

Definitions

Enduring formats: File formats that are expected to be accessible well into the future, like PDF and WAV.

Open formats: File formats that do not require proprietary software to be opened, but rather can be opened and read using free software. The code base is open so that anyone may access it and modify it.

Proprietary formats: File formats that are created by proprietary software that has a closed, private code base and that users must pay for or subscribe to for continued use. Once the software is discontinued, the resulting files cannot be opened or read by other software or operating systems.



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

Enduring/Archival file formats

- Most digital repositories (including language archives) use specific file formats for archiving and preservation.
- Many archives even require depositors to only use specific file formats for their deposits.
- Archives select their recommended or required formats based on a number of criteria:
 - Quality of the format (e.g. WAV vs. MP3)
 - Long-term preservation suitability (“enduring” formats)
 - Openness of the format (open is always preferred over proprietary)
 - Formal standards or de facto standards within the discipline
 - The archival and preservation software and workflows that they use



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

Enduring/Archival file formats

- Recommended/required archival file formats may differ from the formats you use to collect your data, in which case conversions are needed
- “Lossy” compressed formats remove information that is seen as less relevant for human perception, e.g. MP3 audio
- Conversions may result in a loss of quality, e.g. when converting from one “lossy” compressed format to another (e.g. MPEG2->MPEG4), in particular after several conversions
- Converting from a “lossy” compressed format to an uncompressed format does not recover the information that was thrown out (e.g. MP3->WAV)
- Try to collect your materials in archival formats whenever practically possible (e.g. audio recorders can record in uncompressed WAV. Recording uncompressed video is not currently possible with normal camcorders)
- Some archives may allow you to deposit your original files and will convert them for you, others will require you to convert them yourself following their guidance



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

Examples of recommended formats

- **Audio:** uncompressed WAV, 16 or 24 bit, 44.1 kHz or 48 kHz (96 kHz and 192 kHz have no added value for field recordings & result in much larger files)
- **Video:** MPEG4 (However, lots of options still within this format regarding resolution, bit rate, encoding, etc. Check with the archive for recommended settings)
- **Images:** TIFF (for scans or “raw” images) or JPEG (for regular digital camera images)
- Structured multi-tiered text: ELAN “EAF” format, FieldWorks FLEXT (XML-based formats rather than Word or PDF)
- Check with the archive you will be working with for their exact recommendations, requirements and procedures



Day 3

Finish Phase 1: Step 4

Phase 2: Steps 5-6



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

Questions



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

Step 4: Understand Metadata



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

What are metadata?

Metadata

the data
about
your data

Descriptive metadata

contextual data
that can help
users locate the
materials

Structural metadata

data about where an
object is located in a
sequence, hierarchy,
or file structure

Technical metadata

data about the
size, form, &
specifications
of objects

Rights metadata

data about an
object's copyright
status & holder, &
any relevant licenses

Preservation metadata

data used to
assure that a file
has not been
corrupted or lost

Different types of metadata



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

Descriptive Metadata

Examples of *descriptive* metadata for a e.g. video recording

- The **name(s)** of the person or people being recorded;
- The **name** of the person doing the recording;
- The **date** the recording was made;
- The **location** where the recording was made;
- The **language(s)** being spoken or signed;
- The **topic or subject** of the recording.



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

Structural Metadata

- Structural metadata explain relationships between files.
- It is sometimes necessary to incorporate structural metadata into descriptive metadata, depending on the archive's software
- E.g., set of digital photographs of woven designs in fabric that are meant to accompany a PDF document that describes weaving techniques



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

Rights Metadata

Information such as the

- copyright status,
- licensing agreements, and
- restrictions or limitations on access, sharing and use
- traditional protocols for accessing the data

The possibilities for assigning licenses, granting/restricting access, and enforcing traditional protocols vary greatly between archives.



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

Technical Metadata

Technical information about files

- Size
- Format
- Specifications (e.g, recording specs, video codecs)

Crucially important if a repository accepts only certain formats or has file size limits.

Preservation Metadata

Data used to ensure that files are not corrupted or lost.

Workflows that are usually carried out by whomever is responsible for the long-term preservation of the files.



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

Metadata Schema / Standards

- **Metadata Schema:** a standardized system of organizing metadata for cataloging purposes. There are many different metadata schemas, and each one has its own predetermined set of **metadata elements** (or fields, such as author, title, date created, location, etc.) and rules for organizing those elements
- Metadata standards used by language archives include:
 - Dublin Core Metadata Initiative (DCMI, or just DC)
 - Open Language Archives Community metadata (OLAC-Metadata)
 - Component MetaData Infrastructure (CMDI)
 - Metadata Object Description Schema (MODS)
- Familiarise yourself with the metadata requirements of the archive you will be working with



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

Controlled Vocabularies

Controlled Vocabulary (CVs): a fixed set of terms used to describe a given element of metadata; a fixed list of possible values for certain metadata fields.

The purpose of a CV is to provide consistency in the used terminology, thereby improving retrievability

Authority File: Some CV lists are maintained by an authority, such as the Library of Congress or the International Organization for Standardization (ISO), used to disambiguate certain entities, like languages and geographic locations.

You may find that a CV does not provide the exact term you need. Consider using a close match, rather than no value at all. Sparsely filled in metadata reduces the likelihood of your materials showing up in search results.

Sometimes a CV is optional, i.e. one can either select a value from the list or provide a different value. Still consider using a CV value if there is a close match



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

Controlled Vocabulary Ex

Participant Roles

Actor, performer: A person who acts or performs a part in a dramatic production, play, skit, religious pageant.

Analyst: A person who analyzes a recording, text or dataset.

Annotator: A person who annotates a recording, text or dataset.

Author: The writer of a book, article, report, poem, etc.

Collector: A person or organization that was responsible for or that oversaw the collection of the materials contained in an archival collection; the person or organization whose name is on a collection.

Compiler: A person who produces a list or dataset by assembling information or material from other sources.

Consultant: A person who provides expert information on a topic.

Contributor: A person who contributed in some way to creation of the material. Use only if there is not a more specific role.

Creator: The creator of a resource. Use only when nothing more specific is appropriate.

Data technician, keyboarder: A person who entered the data (e.g., into a database program, word processing file, etc.).



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

Language Metadata

Languages in your metadata - include names and authority codes

- **Subject language:** the language(s) that is(are) the target of the project, e.g., Michif
- **Media language:** all language(s) that are heard or seen in the recording or document, e.g., French, Cree, English

Authority Files for languages:

- ISO 639-3 (<https://iso639-3.sil.org/>), e.g., crg (Michif), fra (French), cre (Cree)
- Glottolog (<https://glottolog.org/>), e.g., mich1243 (Michif), stan1290 (French), cree1272 (Cree)

Michif	
	<i>Michif</i>
Native to	Canada
Region	Métis communities in the Prairies; mostly Manitoba, Alberta, Saskatchewan and Northwestern Ontario, Turtle Mountain Indian Reservation in North Dakota
Native speakers	730 (2010 & 2011 censuses) ^[1]
Language family	Mixed Cree–Métis French
Writing system	Latin
Language codes	
ISO 639-3	<input type="text" value="crg"/>
Glottolog	<input type="text" value="mich1243"/>
ELP	Michif
This article contains IPA phonetic symbols. Without proper rendering support, you may see question marks, boxes, or other symbols instead of Unicode characters. For an introductory guide on IPA symbols, see Help:IPA.	



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

Metadata: for whom?

- Nikolaus Himmelmann (2006): “a language documentation is a lasting, **multipurpose** record of a language.” [...] “Users of such a multipurpose documentation would include the speech community itself, national and international agencies concerned with education and language planning, as well as researchers in various disciplines (linguistics, anthropology, oral history, etc.)”
- When compiling metadata, try to think of these different user groups and what would be relevant for them.
- Think about potential (future) users who know nothing about the context of your work, and give them enough information to understand what they've found.



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

General considerations

- Metadata in an archive will be public. Do not write sensitive information in the metadata records, do not include participant's names unless that is what they want, be careful with very precise location information
- Metadata is best collected *during* data collection, i.e. shortly before or after making your recordings, therefore you should make a plan of what metadata you want/need to collect and how you are going to collect it *before* starting your data collection.



PHASE 1
BEFORE

STEP 1
choose your
archive

STEP 2
name your
files

STEP 3
pick enduring
file formats

STEP 4
understand
metadata

Questions about Phase 1



PHASE 2
DURING

STEP 5
collect
metadata

Step 5: Collect Metadata



PHASE 2
DURING

STEP 5
collect
metadata

Collect metadata

- Files need to be described!
- Imagine trying to figure out what was on a set of 20 audio files if you didn't have a description.
- Imagine what would happen if you had thousands of audio files.
- Descriptions need to be made as soon as possible after the recording is made (human memory is short!!)
- Take notes at the time of the recording.
- Keep notes in a structured form (e.g., spreadsheet or metadata tracking software)



PHASE 2
DURING

STEP 5
collect
metadata

Metadata tools

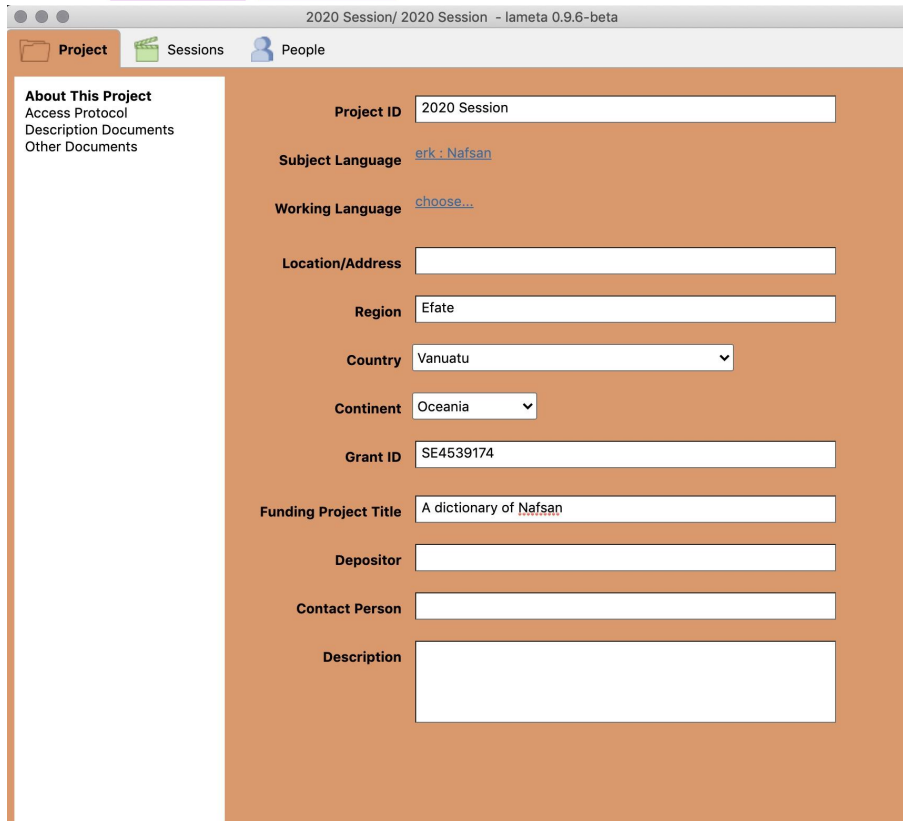
- Paper & pencil
- Software/apps
 - LaMeta
 - Say More
 - Open Data Kit
- Spreadsheets
- Database software



PHASE 2
DURING

STEP 5
collect
metadata

Example: Metadata Software



2020 Session/ 2020 Session - lameta 0.9.6-beta

Project Sessions People

About This Project
Access Protocol
Description Documents
Other Documents

Project ID 2020 Session

Subject Language [erk : Nafsan](#)

Working Language [choose...](#)

Location/Address

Region Efate

Country Vanuatu

Continent Oceania

Grant ID SE4539174

Funding Project Title A dictionary of Nafsan

Depositor

Contact Person

Description

lameta

The Metadata Editor for Transparent Archiving

A system for describing files on your laptop and creating a systematic metadata set that can be exported to different formats suitable for a number of archives

- <https://rebrand.ly/lameta>
- English, Spanish, Portuguese
- Mac & Windows



PHASE 2
DURING

STEP 5
collect
metadata

Example: a metadata spreadsheet supplied by an archive

	Collection ID (e.g. LB8):	Item Identifier (e.g. 1995Elders)	Role: Speaker	Item Title (e.g. Introductory Materials)	Item Description (e.g. Four text stories for interviews)	Content Language (Language as spoken in file)
1						
2	JHER	001	Mona Chuguna	God's Spirit	Acts 1 and 2 with songs	Walmajarri // English
3	JHER	002	Mona Chuguna	Peter and John	Acts 3 and 4 with songs	Walmajarri
4	JHER	003	Mona Chuguna	Stephen	Acts 5 and 6 with songs	Walmajarri // English
5	JHER	004	Mona Chuguna // Peter Skipper	Saul	Acts 8 with songs	Walmajarri // English
6	JHER	005	Mona Chuguna	Peter and Cornelius	Acts 9:32 - 11:18 with songs	Walmajarri
7	JHER	006	Olive Knight // Peter Skipper	Lost Things, The Sower	Luke 15, Mark 4 with songs	Walmajarri
8	JHER	007	Mona Chuguna // Peter Skipper	The Crucifixion	Mark 14:1 - 15:47 with songs	Walmajarri // English
	JHER			Easter	Mark 16:1-8, John 20,21 with songs	Walmajarri // English
9		008	Tommy May // Mona Chuguna			
10	JHER	009	Rela Angie // Mona Chuguna	Jonah	Jonah with songs	Walmajarri // English
	JHER			Christmas, Lazarus	Selections Matthew and Luke, John 11:1-45 with songs	Walmajarri
11		010	Olive Knight // Mona Chuguna			
12	JHER	011	Tommy May	Elijah and Ahab	1Kings 17, 18 with songs	Walmajarri // English
13	JHER	012	Olive Knight	Letter to Timothy	1Timothy 1:1 - 4:5 with songs	Walmajarri // English
14	JHER	013	Rena Pindan	Abraham 1	Genesis 15 - 19 with songs	Walmajarri
	JHER			Abraham 2	Genesis 20,, 21, 22, 24 with songs	Walmajarri
15		014	Peter Skipper			



PHASE 2
DURING

STEP 5
collect
metadata

What metadata? Descriptive

- The participants/contributors (note everyone involved in the process or event)
- Date and location
- Languages that are spoken in each recording, (**especially important in a multilingual environment)
- The context and content of the file.
Why is the recording being made?
- **Create meaningful (event) titles in the subject language, with translation in media language(s)**



PHASE 2
DURING

STEP 5
collect
metadata

What metadata? Rights

- Did you get **informed consent** for recording?
- What **rights** apply to the event / recording / document / image?
- What **protocols** apply to the event / recording / document / image?
- Who is the rights holder? (Tribe, individual(s), non-profit, etc.)
- Is there a Memorandum of Understanding (MOU) or other agreement?
- Information/metadata about the participants
 - Do they want to be named or remain anonymous?
 - What other info do you need/plan to collect about people?
 - Tribe, clan, age, gender identity, languages spoken, etc.



Day 4

Finish Phase 2: Step 6

Phase 3: Steps 7-9



PHASE 2
DURING

STEP 5
collect
metadata

STEP 6
evaluate
your files

Step 6: Evaluate Your Files



PHASE 2
DURING

STEP 5
collect
metadata

STEP 6
evaluate
your files

Materials Appraisal

General principle: a deposit in an archive is not a private dropbox which one can dump stuff and sort it out later....

Analogy:

- **Collection = Publication**
- **Curating & appraising = Editing**

Consider:

- **Quality** (sound, image)
- **Uniqueness** (only recording of X)
- **Relevance** (to the language or culture)
- **Sensitivity** (protocols, access, rights)



PHASE 2
DURING

STEP 5
collect
metadata

STEP 6
evaluate
your files

Quality: audio / video / images

Poor sound quality:

- no sound
- some sound
- noisy sound

Poor video/image quality:

- too light
- too dark
- blurry
- wrong people in frame
- Embarrassing shots, scenes (consider the dignity of language users!)

versus

***Uniqueness of the
recording/image***



Could you record again?



PHASE 2
DURING

STEP 5
collect
metadata

STEP 6
evaluate
your files

Uniqueness



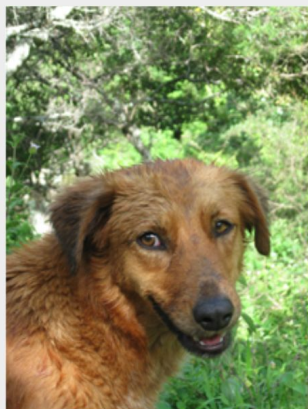
Near duplicate photos from McIntosh (2010), used with permission

PHASE 2
DURING

STEP 5
collect
metadata

STEP 6
evaluate
your files

Relevance to Language/Culture



Photographs of dogs and cats from Cruz, Cruz Baltazar & Cruz Cruz (2009), used with permission.



PHASE 2
DURING

STEP 5
collect
metadata

STEP 6
evaluate
your files

Sensitivities

- Informed consent / permissions - iterative process
- Cultural protocols / traditional knowledge / TK labels
- Private information (obvious or buried deep in recordings)
- Dignity for everyone involved
- Politically sensitive content
- Socially sensitive content
- Consider recordings and transcription/translation



PHASE 2
DURING

STEP 5
collect
metadata

STEP 6
evaluate
your files

Questions about Phase 2



PHASE 3
AFTER

STEP 7
prepare
your deposit

Step 7: Prepare Your Deposit



PHASE 3
AFTER

STEP 7
prepare
your deposit

Primary activities in preparing an archival deposit

- Understanding archives' requirements on file structure
- Developing strategies for arranging files into folders
- Enhancing discoverability with informative titles
- Converting files to required formats



PHASE 3
AFTER

STEP 7
prepare
your deposit

File Arrangement

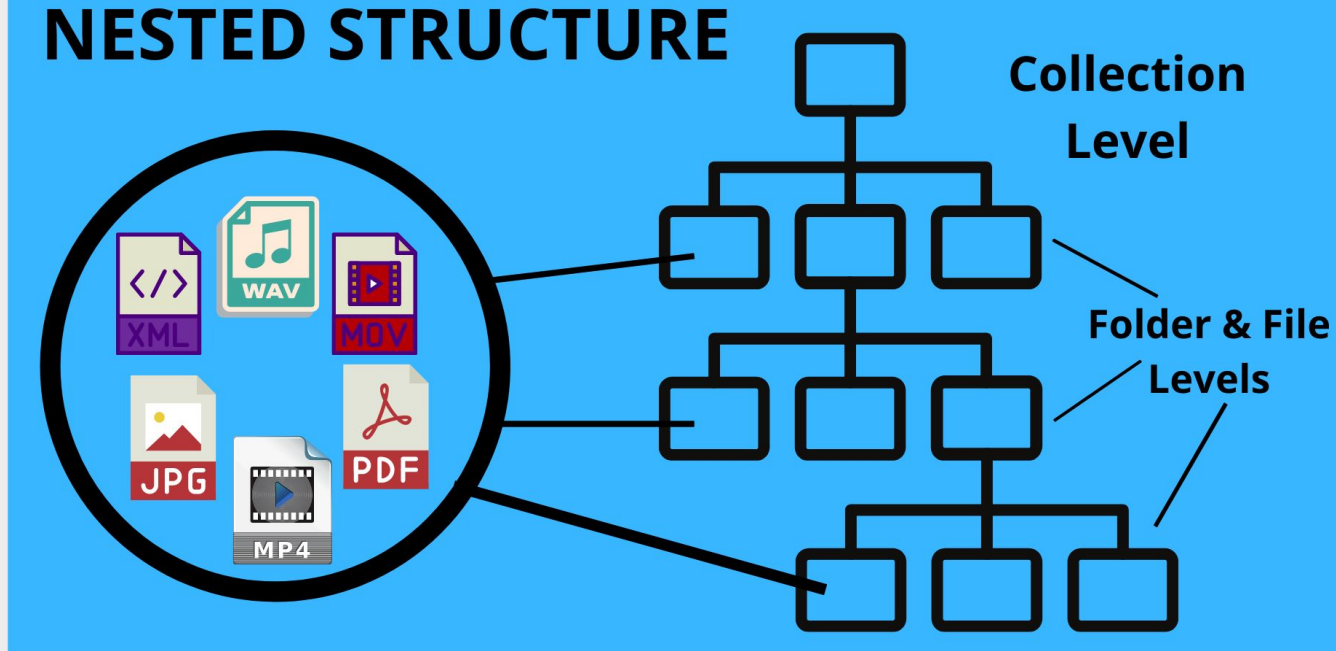
- Refers to how digital files are organized, ordered, and grouped together in a digital environment, e.g., on a personal hard drive or in a digital archive.
- Most digital archives have 3 units/levels of organization:
 - **Collections**, which contain ...
 - **Folders**, which contain ...
 - **Media files**
- Each unit of organization requires its own metadata.



PHASE 3
AFTER

STEP 7
prepare
your deposit

Nested file structure



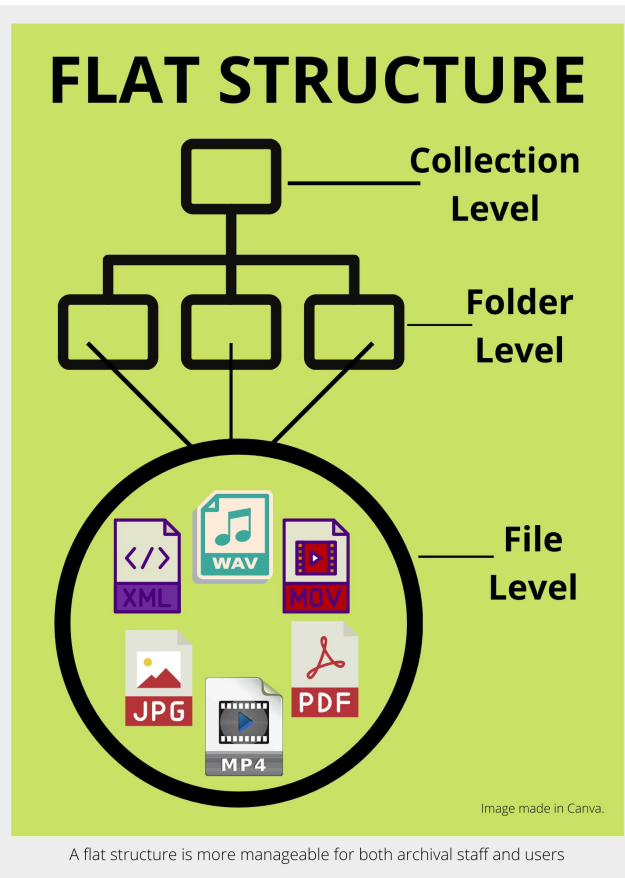
A nested file structure can be challenging for both the user and the archive to navigate and manage



PHASE 3
AFTER

STEP 7
prepare
your deposit

Flat file structure



Arrangement strategies within a flat structure

An **arrangement strategy** is the strategy or process used to organize files into cohesive units for archiving, storage and preservation.

- **recording session** - commonly used in language archives
- language variety
- location
- Speaker
- date
- media format
- experimental protocols
- questionnaires or surveys
- replication data, e.g., to accompany a publication

Planning to archive progressively may
affect arrangement!



PHASE 3
AFTER

STEP 7
prepare
your deposit

Discoverability

Discoverability is the ease with which users can find your collections and the items within them and understand the connections between files.

Improve discoverability by

- Including detailed metadata at all levels of the archive's structure (collection, folder, and media file).
- Writing clear, concise and relevant titles in both the Indigenous language and the language of the archive.
- Writing titles that can be easily understood by archive users (e.g., avoid opaque abbreviations)



PHASE 3
AFTER

STEP 7
prepare
your deposit

Discoverability by title

Avoid redundant titles for items when possible...

12. [Elicitation session](#) (11 Sep 2014) (3 digital files, with audio) 🗣️▶️
13. [Elicitation session](#) (18 Sep 2014) (4 digital files, with audio) 🗣️▶️
14. [Elicitation session](#) (20 Sep 2014) (3 digital files, with audio) 🗣️▶️
15. [Elicitation session](#) (20 Sep 2014) (3 digital files, with audio) 🗣️▶️
16. [Elicitation session](#) (22 Sep 2014) (2 digital files, with audio) 🗣️▶️
17. [Elicitation session](#) (22 Sep 2014) (2 digital files, with audio) 🗣️▶️

Hector Zapana Almanza, Kenneth Baclawski, Spencer Lamoureux, Herman Leung, Lev Michael, Zachary O'Hagan, Alfonso Otaegui, Nicholas Rolle, Kamala Russell, Hannah Sande, Eva Schinzel, and Amalia Horan Skilton. **Berkeley Field Methods: Aymara, 2014-10**, Survey of California and Other Indian Languages, University of California, Berkeley, <http://dx.doi.org/doi:10.7297/X2S18oHS>



Discoverability by title

Disambiguate titles as much as possible

5. [Small-group elicitation session on possessive pronouns] (09 Sep 2015) (4 digital files, with audio) 🎧▶
6. [Small-group elicitation session on demonstratives] (09 Sep 2015) (1 digital file, with audio) 🎧▶
7. [Small-group elicitation session on possessive pronouns] (15 Sep 2015) (2 digital files, with audio) 🎧▶
8. [Small-group elicitation session on adjectives] (15 Sep 2015) (2 digital files, with audio) 🎧▶
9. [Small-group elicitation session on plural nouns] (15 Sep 2015) (1 digital file, with audio) 🎧▶
10. [Small-group elicitation session on demonstratives] (16 Sep 2015) (1 digital file, with audio) 🎧▶

Guy Tchatchouang, Geoff Bacon, Steven Bird, Andrew Cheng, Emily Clem, Virginia Dawson, Larry M. Hyman, Anna Jurgensen, Erik Hans Maier, and Alice Shen. **Berkeley Field Methods: Tswefap, 2020-11**, Survey of California and Other Indian Languages, University of California Berkeley, <http://dx.doi.org/doi:10.7297/X2No152S>



PHASE 3
AFTER

STEP 7
prepare
your deposit

Converting file formats

Some files may need to be converted to other formats when it comes time to archive

- audio
- video
- text
- photographs
- databases and tabular data
- zipped files

Check with the archive for its requirements!



PHASE 3
AFTER

STEP 7
prepare
your deposit

Self-Archiving

Also called *self-depositing*, *pre-archiving*, *self-upload*

General steps:

- Submit paperwork (license or other agreements) & establish your account
- Add descriptive metadata
- Upload digital files

Details vary by archive!

Plan for it to take a lot of time!

[CLA PREARCHIVE](#) [[Help guide](#)]

Collection 2014-13

[Create a new file bundle](#) | [Set up Contributions, Languages, or Places for this Collection](#)

Title: Caquite Field Materials

Associated materials: ?

Materials dating from 2014 to 2018 are additionally archived with Endangered Languages Archive (ELAR) at SOAS, London, and available online here: <https://elar.soas.ac.uk/Collection/MP11032021>.

Historical information: ?

As of February 2020, Caquite is a vital Nijagantsi (aka Kampa) Arawak language of southeastern Peru with 300-400 speakers spread across 7 communities located in the headwaters that feed the Tambo and Urubamba rivers in the regions of Junín and Cusco, respectively: Tsoroja, San Luis de Korinto, Taini, Kitepampani, Dios Maseca, Mashía, and Mankoriari. Other Nijagantsi languages include Ashaninka, Asheninka, Matsigenka, Nanti, and Nomatsigenka. These languages boast some of the largest speech communities of lowland Amazonia, with many tens of thousands of speakers in total. Caquintes first entered into sustained contact with non-indigenous outsiders in 1976, although they had intermittent contact both peaceful and not with

Scope and content: ?

Audio and video recordings of elicitation sessions and texts; field notes; derivative materials

[Update descriptive metadata](#)



PHASE 3
AFTER

STEP 7
prepare
your deposit

Step 7 Questions



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

Step 8: Practice Progressive Archiving



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

Progressive Archiving

Manioc Preparation

Primary Data

- manioc.mov
- manioc-01.jpg
- manioc-02.jpg
- manioc-03.jpg

STEP 1

Secondary Data

- manioc-transcription.eaf
- manioc-translation.eaf

STEP 2

Publication

- manioc-prep-guide.pdf

STEP 3

- Portions of a collection are archived in stages, rather than all at once at the end of a career.
- Primary recordings are added to the archive immediately or shortly after they are collected.
- Secondary materials and analyses are added later, as they are finalized.
- Interactive process.
- If recording / materials creation is ongoing, then archiving should be done at regular intervals or in increments.
- A.k.a. *incremental archiving*



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

Archived Primary Materials

Manioc Preparation

Primary Data

- manioc.mov
- manioc-01.jpg
- manioc-02.jpg
- manioc-03.jpg

STEP 1

Secondary Data

- manioc-transcription.eaf
- manioc-translation.eaf

STEP 2

Publication

- manioc-prep-guide.pdf

STEP 3



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

Added Secondary Data

Manioc Preparation

Primary Data

- manioc.mov
- manioc-01.jpg
- manioc-02.jpg
- manioc-03.jpg

STEP 1

Secondary Data

- manioc-transcription.eaf
- manioc-translation.eaf

STEP 2

Publication

- manioc-prep-guide.pdf

STEP 3



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

Added Publication

Manioc Preparation

Primary Data

- manioc.mov
- manioc-01.jpg
- manioc-02.jpg
- manioc-03.jpg

STEP 1

Secondary Data

- manioc-transcription.eaf
- manioc-translation.eaf

STEP 2

Publication

- manioc-prep-guide.pdf

STEP 3



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

Technical issues in adding to archival collections

- All archives allow you to **add folders** to collections
- Not all archives allow you to **add files** to folders
- Files cannot usually be deleted
- New **versions of files** must be added and related to existing files
- Archives differ in how they handle versions

	Innate versioning	Sequential version numbers	Meaningful suffixes
2018 video	2018-toro.mp4	2018-toro.mp4	2018-toro.mp4
2018 transcription	2018-toro.eaf	2018-toro_v01.eaf	2018-toro_transcription.eaf
2019 transcription & translation	2018-toro.eaf	2018-toro_v02.eaf	2018-toro_translation.eaf

File versioning methods



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

Avoid deleting archived materials!!

3 important reasons to avoid deleting archived materials

1. If a file has ever been publicly available for viewing or use, it might have been cited in a publication. If a reader of the publication goes to the archive to try to find it, the record must be there. Anything used as "data" must be citable and accessible in order to be verifiable and/or replicable.
2. Digital preservation involves multiple backups on various kinds of media. Thus, if a file is deleted from an archive, it is not easy or even completely feasible to delete every copy of it.
3. If the file was previously downloaded by a user (prior to deletion), that user still has a copy the file and might use it and cite it. (the file is in the digital wild)

Thus, be sure to archive only what you are certain will be appropriate to share.



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

Benefits of archiving often

- Benefits from results
 - Dissemination of materials to community members and other stakeholders on a regular schedule
 - Archived materials can feed back into subsequent ILR efforts
 - Satisfies expectations/requirements of funders, committees, etc.
- Benefits to archiving process
 - Avoids overwhelming scale of archiving at the end of a project or career
 - Data management is easier when done in small batches
 - Context is recent and remembered more easily (e.g., for metadata)
 - Communities can be engaged more readily in decision-making (e.g., regarding access)
 - Safeguards against inadvertent loss or obsolescence of materials



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

STEP 9
describe your
collection

Step 9. Describe Your Collection



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

STEP 9
describe your
collection

Types of collection descriptions

Collection description: basic contextual information that explains the collection in broad terms to anyone who is unfamiliar with you, your project, your research, and/or the language(s) that are documented in the collection. Each time you add more materials to the archived collection, you should update the collection description to include those materials, the dates they were collected, new funding streams, and other relevant new information.

Collection guide: A stand-alone guide to a collection summarizes the entirety of a collection in a single document. It serves to explain in detail the history and context of the documentation efforts, the people involved, the time frame, the funding streams, orthographic conventions, abbreviations, and any additional information that might be necessary for the collection to be accurately understood and reused. If this guide is accompanied by a comprehensive list of the collection's contents, it can serve as a *finding aid* that can direct users to particular materials within the collection.

Finding aid: a collection guide that includes a comprehensive list of the collection's contents; they can reflect multiple languages, multiple arrangements.



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

STEP 9
describe your
collection

Minimum Information

Should include:

- Reason for collection
- History of the project/materials
- Overview of the contents
- Language(s) represented
- Community(s) represented
- Collection title/identifiers



Additional Information

Can include:

- Access statement
- Relevant dates
- Relevant funding
- Relevant individuals
 - Biographical sketches
- An inventory of the collection materials (finding aid/collection guide)
 - description, hierarchical structure, conventions
- Citation information
- Name/location of repository(s) (if not obvious)
- Tech specs
- Keywords/genres/subjects
- Expected additions
- Associated publications

PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

STEP 9
describe your
collection

Additional Information

Can include:

- Access statement
- Relevant dates
- Relevant funding
- Relevant individuals (Biographical sketches)
- An inventory of the collection materials
- Citation information
- Name/location of repository(s) (if not obvious)
- Tech specs
- Keywords/genres/subjects
- Expected additions
- Associated publications



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

STEP 9
describe your
collection

Some examples in archives

- **APS:** <https://indigenousguide.amphilsoc.org/>
- **ELAR:** <https://elararchive.org/collections/>
- **NAL:**
<https://samnoblemuseum.ou.edu/collections-and-research/native-american-languages/native-american-languages-collections/>
- **CLA:** <https://cla.berkeley.edu/browse-collections.php>
- **AILLA:** https://ailla.utexas.org/islandora/object/ailla%3Acollection_collection
- **PARADISEC:** <https://catalog.paradisec.org.au/collections/search>



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

STEP 9
describe your
collection

Sample collection descriptions in LD&C

Gawne, Lauren. 2018. A Guide to the Syuba (Kagate) Language Documentation Corpus. Language Documentation & Conservation 12. 204-234. <http://hdl.handle.net/10125/24768>

Franjeh, Michael. 2019. The languages of northern Ambrym, Vanuatu: A guide to the deposited materials in ELAR. Language Documentation & Conservation 13: 83-111. <http://hdl.handle.net/10125/24849>

Caballero, Gabriela. 2017. Choguita Rarámuri (Tarahumara) language description and documentation: a guide to the deposited collection and associated materials. Language Documentation & Conservation 11: 224-255. <http://hdl.handle.net/10125/24734>

Salffner, Sophie. 2015. A guide to the Ikaan language and culture documentation. Language Documentation & Conservation 9. 237-267. <http://hdl.handle.net/10125/24639>

Oez, Mikael. 2018. A Guide to the Documentation of the Beth Qustan Dialect of the Central Neo-Aramaic Language Turoyo. Language Documentation & Conservation 12. 339-358. <http://hdl.handle.net/10125/24773>



PHASE 3
AFTER

STEP 7
prepare
your deposit

STEP 8
practice
progressive
archiving

STEP 9
describe your
collection

Questions about Phase 3



In Closing

- Don't be intimidated! Not as scary/complicated as this might sound
- Some work on the front end saves A LOT of work later
- Archives are long-term, stable solutions for preservation & access
- Everyone's goal is to make language information safe and accessible
- Archiving is part of our responsibility to our people, friends, elders and descendents
- **Don't let the perfect be the enemy of the good: it's better to archive something that is imperfect or incomplete than to archive nothing at all!**



References & Acknowledgement

The contents of this workshop come directly from

Kung, Susan, Ryan Sullivant, Elena Pojman & Alicia Niwagaba. 2020. *Archiving for the Future: Simple Steps for Archiving Language Documentation Collections* [OER], <https://archivingforthefuture.teachable.com/>. Licensed under a [CC BY-SA 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).

Kung, Susan S., Paul Trilsbeek, Mandana Seyfeddinipur, Nick Thieberger, Zachary O'Hagan, Raina Heaton. 2021. Relating the Past, Present & Future: Archiving Language Collections. Workshop presented at the *7th International Conference on Language Documentation & Conservation*, University of Hawai'i at Manoa, March 4-7, 2021. <http://bit.ly/ICLDC7ARCHIVING>. Licensed under a [CC BY-SA 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).

The OER *Archiving for the Future: Simple Steps for Archiving Language Documentation Collections* is based upon work supported by the National Science Foundation under Grant No. BCS-1653380. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.



Cite this workshop:

Kung, Susan S. and Jaime Pérez González. 2022. Archiving for the Future [workshop slides]. CoLang 2022. [CC BY-SA 4.0 International](https://creativecommons.org/licenses/by-sa/4.0/).

