

Using language archives: finding and depositing resources of minoritised languages.

FOSTERLANG workshop, Donostia, 28 May 2026

Paul Trilsbeek
The Language Archive
Max Planck Institute for Psycholinguistics

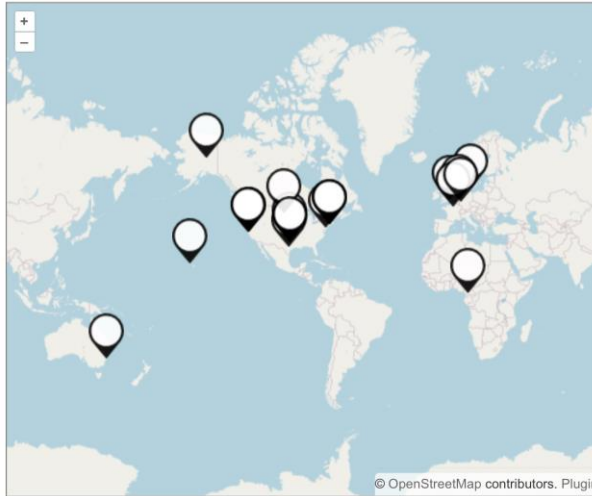
Finding Language Resources

Finding language resources

- Web search engine (Google, Bing, DuckDuckGo, etc.)
- AI Chatbot (ChatGPT, Claude, Gemini, etc.)
- Individual language archive website/catalogue (e.g. via DELAMAN website)
- “Metadata aggregators”
 - CLARIN Virtual Language Observatory (VLO)
 - OLAC search engine

DELAMAN

- Digital Endangered Languages and Musics Archives Network (DELAMAN) is an umbrella body for archives that preserve materials on endangered languages and cultures



- Currently 14 full members and 5 associate members
- Some archives with a regional focus, some with more global coverage
- Descriptions and links for member archives can be found on the website

www.delaman.org

The CLARIN Infrastructure

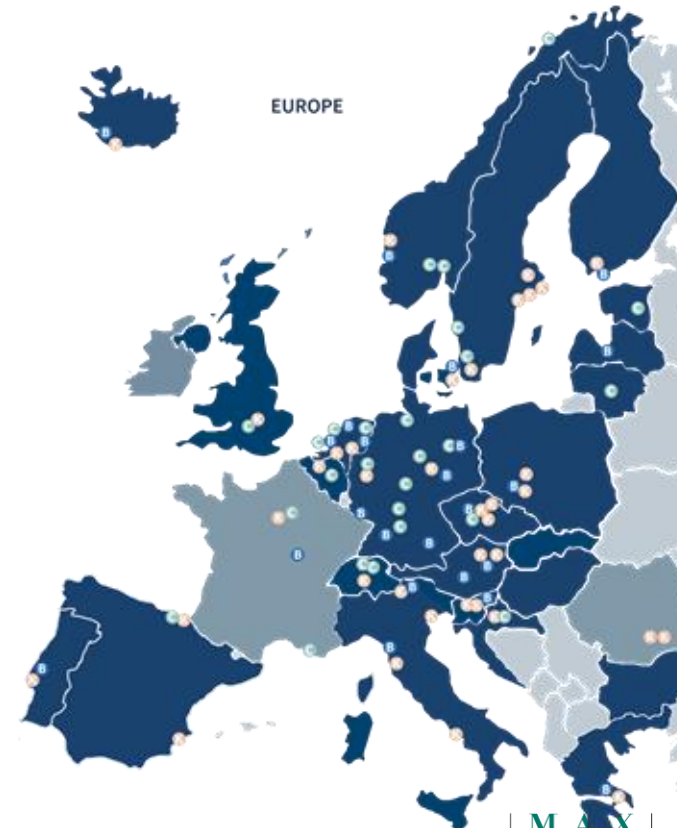
“CLARIN - or Common Language Resources and Technology Infrastructure - is a digital infrastructure which provides easy and sustainable access to a broad range of language data and tools to support research in the humanities and social sciences, and beyond.”

- European Research Infrastructure Consortium (ERIC) legal structure, funded by contributions from the member countries
- Participating institutions from 24 member countries, 3 “observer” countries

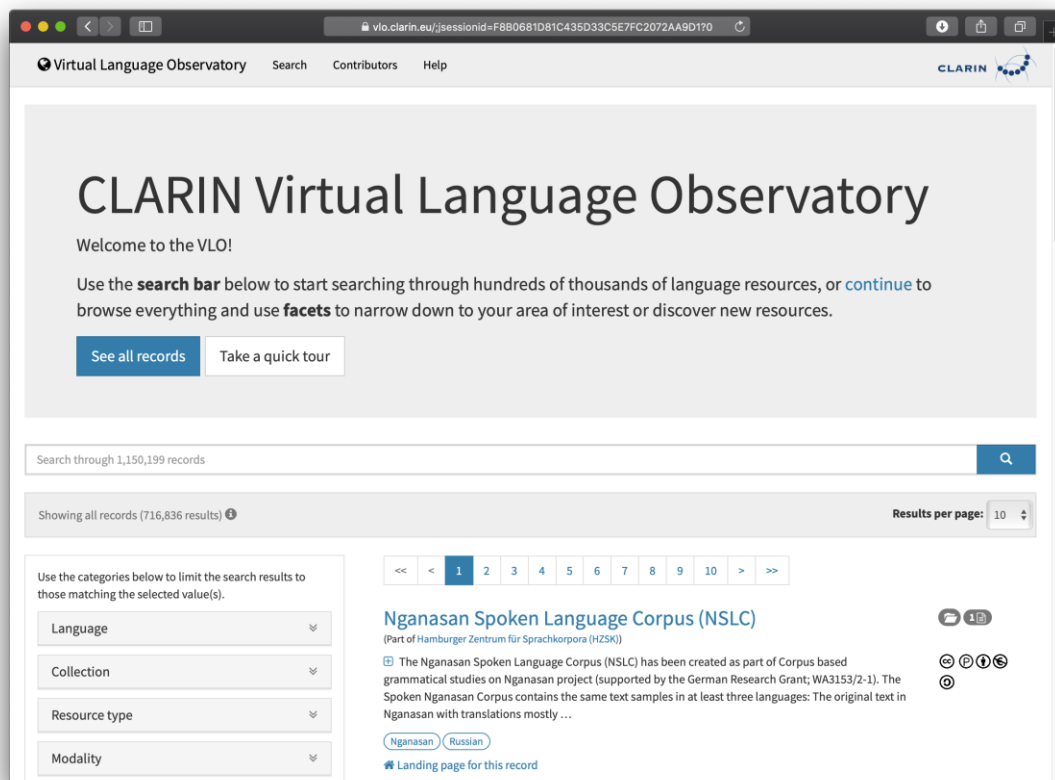
www.clarin.eu

CLARIN Centres

- Centres within the CLARIN infrastructure offer the scientific community access to resources, services and knowledge on a sustainable basis
- 25 certified B-Centres (Service Providing Centres), with different foci, e.g. regional focus or focus on a certain linguistic sub-discipline
- Many of them have a repository for archiving and sharing language data



CLARIN Virtual Language Observatory (VLO)



Virtual Language Observatory Search Contributors Help

CLARIN

CLARIN Virtual Language Observatory

Welcome to the VLO!

Use the **search bar** below to start searching through hundreds of thousands of language resources, or [continue](#) to browse everything and use **facets** to narrow down to your area of interest or discover new resources.

[See all records](#) [Take a quick tour](#)

Search through 1,150,199 records

Showing all records (716,836 results) Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

- Language
- Collection
- Resource type
- Modality

<< < 1 2 3 4 5 6 7 8 9 10 > >>

Nnganasan Spoken Language Corpus (NSLC)

(Part of Hamburger Zentrum für Sprachkorpora (HZSK))

The Nnganasan Spoken Language Corpus (NSLC) has been created as part of Corpus based grammatical studies on Nnganasan project (supported by the German Research Grant; WA3153/2-1). The Spoken Nnganasan Corpus contains the same text samples in at least three languages: The original text in Nnganasan with translations mostly ...


[Nnganasan](#) [Russian](#)

[Landing page for this record](#)

CLARIN Virtual Language Observatory (VLO)

vlo.clarin.eu

Open Language Archives Community (OLAC)



OLAC: Open Language Archives Community

SEARCH THIS SITE:

[HOME](#) | [DOCUMENTS](#) | [ABOUT](#) | [ARCHIVES](#)
[NEWS](#) | [ORGANIZATION](#) | [TOOLS](#) | [SERVICES](#)

OLAC Mission

OLAC, the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources.

News

OLAC Joins the Linguistic Linked Open Data Cloud: The OLAC system has now been integrated with LLOD... [More...](#)

New OLAC Page Listing Archive Submission Policies: As a service to linguists who are in search of an archive that could receive a deposit... [More...](#)

New OLAC Search Service: In December 2010, to mark our 10th anniversary, OLAC announces a new search service... [More...](#)

[More news ...](#)

Documents

[OLAC Standards](#) - specify how OLAC operates


[Recommendations](#) - express consensus of OLAC members regarding language resource archiving

[Notes](#) - background information and guidance for implementers

General Information:
[Overview](#) | [FAQ](#) | [Implementers' FAQ](#)

Find Language Resources

Use OLAC's search engine at: <http://search.language-archives.org/>



Search

Archive: -- all archives --

Region: [Africa](#) [Americas](#) [Asia](#) [Europe](#) [Pacific](#)

OLAC Coverage

OLAC Archives contain over 300,000 records, covering resources in half of the world's living languages. [More statistics on coverage.](#)

Join the OLAC Community

Sign-up for the [OLAC mailing list](#) and stay current with standards and best practices for language resource archiving ([Archives](#)).

The OLAC coordinators may be contacted [via email](#).

OLAC metadata search (preliminary version, new launch planned for September)

The screenshot shows a web browser window with the URL `search.language-archives.org/?size=n_5_n&filters%5B0%5D%5Bfield%5D=dc.format.txt.keyword&filt...`. The page title is "Open Language Archives Search". Below the title are links for "Imprint" and "Privacy". A search bar contains the text "Search" and a blue "Search" button. On the left side, there are filter sections:

- CLEAR FILTERS**
- OLAC LANGUAGE (THE MATERIAL)**: A search box contains "eus", with a dropdown menu showing "eus" (30).
- LINGUISTIC FIELD**: A search box contains "Filter Linguistic field", with a dropdown menu showing "sociolinguistics" (3).
- OLAC LANGUAGE (THE RESOURCE IS ABOUT)**: A search box contains "Filter Olac language (", with a dropdown menu showing "eus" (46) selected.

On the right side, the search results are displayed. It shows "Showing 1 - 5 out of 46" and a "Show 5" dropdown. The first result is:

- CCausality Across Languages (CAL): Causality in Discourse, basque**
- Description**: Causality in Discourse: This subproject investigates the pragmatic principles which govern the representation of causality in discourse, with a focus on the narrative genre. Content of each session: -Demographic data -Video recording -Transcription:
- Contributor**: María 023
- Publisher**: María Louro Mendiguren Universidade de Santiago de Compostela
- License**

Open Language Archives Community (OLAC)

www.language-archives.org

Language codes

Language codes

- Many languages go by a number of different names, including different names in different languages
- Some different languages have the same name
- In order to have a unique identifier for a given language, language "codes" have been developed
- ISO 639 family: official standard from the International Organization for Standardization
- Glottocodes: codes given to each "Langoid" (Language, dialect, family) in the Glottolog database (glottolog.org)

ISO 639

Set (past Part) ⇄	Former name (<i>Codes for the representation of names of languages – ...</i>) ⇄	Language Coding Agency (formerly registration authority) ⇄	First edition ⇄	Current ⇄	No. in list (as of 12 July 2023) ⇄
Set 1	<i>Part 1: Alpha-2 code</i>	Infoterm	1967 (original ISO/R 639)	2023	183
Set 2	<i>Part 2: Alpha-3 code</i>	Library of Congress	1998	2023	482 + 20 B-only + 4 special + 520 for local use ^{[5][6]}
Set 3	<i>Part 3: Alpha-3 code for comprehensive coverage of languages</i>	SIL International	2007	2023	7,916 + 4 special + 520 for local use ^[7]
(ISO 639-4)	<i>Part 4: Implementation guidelines and general principles for language coding</i>	ISO/TC 37/SC 2 ↗	2010-07-16	2023	(not a list)
Set 5	<i>Part 5: Alpha-3 code for language families and groups</i>	Library of Congress	2008-05-15	2023	115 (including 36 remainder + 29 regular groups from ISO 639-2) ^[8]
(ISO 639-6)	<i>Part 6: Alpha-4 representation for comprehensive coverage of language variants (withdrawn)</i>	Geolang	2009-11-17	withdrawn	21,000+

Source: [wikipedia.org](https://en.wikipedia.org/wiki/ISO_639)

ISO 639 Set 3

- 3-letter codes, aims to cover all known natural languages
- First version primarily based on language codes used in the *Ethnologue: Languages of the World* publication by SIL International
- Registration authority is SIL International. Change requests are submitted to SIL and evaluated by their staff.
- Widely used in language archives
- Provides links to [Ethnologue](#), [Glottolog](#) and Wikipedia
- Some issues: little information available about the languages, information sometimes incorrect, objections against having a christian faith-based organisation being in charge of the standard (further reading: https://en.wikipedia.org/wiki/ISO_639-3)

Glottolog

- Online bibliographic database of the world's lesser-known languages
- Created by Sebastian Nordhoff and Harald Hammarström at the MPI for Evolutionary Anthropology in Leipzig
- Curated by Harald Hammarström, Martin Haspelmath et al., change requests via [GitHub](#)
- Shows language family classification
- Aims to only show languages that exist and are distinct, as validated by the editors
- Shows (and lets you search by) alternative names as found in different sources
- Provides links to e.g. [ISO 639-3](#), [Ethnologue](#), [Endangered Languages Project](#), [WALS](#), Wikidata, Wikipedia

Potential issues when obtaining or working
with language data from archives

Access policies

- Many language materials are not freely downloadable
- Access policies vary per archive and often per collection, sometimes also within collections
- E.g. The Language Archive distinguishes 4 access levels, which can be applied to different parts of a collection or even to individual files
- Procedures to obtain access vary, sometimes an explicit written request is needed

Open

Registered

Academic

Restricted

Issues with access requests

- For some archives, the archive depends on a decision from the depositor before access can be granted to many collections
- Understandable given privacy considerations, especially with people from indigenous communities or other minority groups, however:
- Currently a barrier for people to get access, reluctance to request access in the first place and processing of requests may take a long time
- Not a sustainable solution if the depositor can no longer be contacted (retirement, death...)

Issues with content

- Content has moved or is no longer available
- Interface and/or metadata only in English
- Incomplete metadata
- Recording quality of media not ideal (field situation is not a recording studio)
- Media not or only partially transcribed/annotated (hugely time-consuming work)

Issues with content

- Files may require conversions before you can work with them in your environment
- Even though file formats might be (de-facto) standards, many degrees of freedom within the formats for specific conventions.
- Transcription/annotation/glossing conventions not always documented

Please Cite!

Please **cite** language resources if you use them for your research! (using DOI/Handle/other Persistent Identifier if available)

Important to acknowledge those who created the materials as well as the repository that hosts them. Hopefully, in the long run this will lead to better academic recognition for creating and archiving datasets.

General archiving and long-term preservation principles

Archiving and long-term preservation

- Archiving and sharing of research data part of normal scientific practice nowadays (reproducibility/verifiability, re-use)
- Preservation of valuable (cultural) resources for current and future generations
- A lot of focus on “open” science, [FAIR](#) data (Findable, Accessible, Interoperable, Reusable)
- Truly “open” data in the sense of unrestricted access obviously not possible for sensitive materials. [CARE](#) Principles for Indigenous Data Governance
- Lots of research data repositories exist (3155 listed on [re3data.org](#))
- Generalist vs. Domain repositories: in general, best to use a certified* domain repository as it can better cater for the needs of the particular research community.

* According to [CoreTrustSeal](#), [Nestor Seal / DIN 31644](#) or [ISO 16363](#)

Archiving and long-term preservation

- Digitisation of analogue media (tape, film, etc.) before they deteriorate
- Digital preservation:
 - Preservation of “bit-streams”: making sure that stored digital objects (files) on a storage medium are preserved -> backups, replication to other locations, timely migration of storage media
 - Preservation of content: making sure that the content of the objects remains interpretable over time -> file format migration before they become obsolete (ideally without loss of information), use of file formats that are well suited for preservation (preferably open standards)
- Good quality Metadata is essential for current and future use of the materials
- Further aspects of “Trusted Digital Repositories”: 16 requirements of the CoreTrustSeal www.coretrustseal.org

Preparing materials for depositing with an archive

Contact the archive

- Look e.g. on the DELAMAN website for an archive that is appropriate for the language materials you would like to deposit (geographic focus)
- See whether they have published a policy on their website for accepting deposits
- Contact the archive with your request
- If the archive accepts your deposit, familiarise yourself with the archive's workflows and requirements
- Some archives offer a self-deposit web interface, others do not

Preparing materials for deposit

- Prepare files in appropriate formats (recommended or required by the archive)
 - E.g. the archive may require uncompressed audio files in WAV format rather than MP3, or documents in PDF format rather than Microsoft Word.
- Name files in a consistent and useful manner
 - E.g. rather than “IMG_1816.MP4”, perhaps something like “2026-05-28_Donostia_workshop_3.mp4”
- Prepare metadata in a format that is recommended or required by the archive. Know which fields are needed and collect the metadata systematically, ideally during the data collection

Access Policies

- If the archive offers different access levels, think about which level is most appropriate for your data, in line with consent that speakers have given

Metadata for language resources

Metadata for language resources

- Commonly used standards:
 - **OLAC**: Developed by the Open Languages Archiving Community, extension of Dublin Core metadata. Limited number of elements.
 - **CMDI**: Developed within the CLARIN community, not one single schema but a framework for defining "profiles" composed out of "component" building blocks that can be shared.
 - TEI: header of the Text Encoding Initiative, different profiles exist for different types of resources, e.g. "Transcription of Spoken Language" TEI-based standard (ISO 2462:2016)

CMDI metadata

- Component MetaData Infrastructure, developed within CLARIN
- Not a single schema for metadata, but a framework that enables the composition of metadata “profiles” from building blocks that are stored in a central registry
- Enables the creation of metadata profiles that have been tailored for specific types of data, while still ensuring a level of interoperability
- Many CMDI profiles have been created. All are stored in the CLARIN CMDI Component Registry.
- Most CLARIN Centres with a repository require the use of specific CMDI profiles -> Familiarise yourself with the metadata requirements of the repository you will be working with (not just in the case of CLARIN)

OLAC metadata

- Developed by the Open Language Archives Community (OLAC)
- Fixed schema, "Qualified Dublin Core" i.e. refinement of the basic Dublin Core standard for use with language resources
- Limited number of fields (pro and con)
- Easily interoperable with tools/frameworks that work with Dublin Core

www.language-archives.org/OLAC/metadata.html

Pbh86_Yor05b_Ancha1

Object Type: Folder
In Folder: **panare-caceres-0628**

Browse Archive > panare-caceres-0628 > Pbh86_Yor05b_Ancha1

Sort by type ▲

<p>Pbh86_Yor05b_Ancha1 Pbh86_Yor05b_Ancha1</p> <p>Session Information Title: Ancha (1), the shaman's anteat... Description: [This text was told as p... Date created: 1986-03-30</p> <p>Location Continent: Americas Country: Venezuela Region: Bolívar Address: Caño Amarillo</p> <p>Project ID: MDP0407 Name: 0628-MDP0407 Description: The Panare documenta...</p> <p>Topics Topic: Dangers</p> <p>Keywords Keyword: Shamanic beings</p> <p>Content</p>	<p>Genre: Interactive discourse</p> <p>Languages</p> <p>Language Name: Panare Description: Content Language</p> <p>Language Name: Spanish Description: Working Language</p> <p>Language Name: English Description: Working Language</p> <p>Actor Full Name: Marie Claude Mattei MuL Role: recorder</p> <p>Actor Full Name: Marie Claude Mattei MuL Role: researcher</p> <p>Actor Full Name: Marie Claude Mattei MuL Role: transcriber</p>	<p>Actor Full Name: Marie Claude Mattei MuL Role: translator</p> <p>Actor Full Name: Uñe'; Felipe Argoto Role: speaker</p> <p>Actor Full Name: Rafael Moncada Role: interviewer</p> <p>Actor Full Name: Rafael Moncada Role: transcriber</p> <p>Actor Full Name: Rafael Moncada Role: translator</p> <p>Actor Full Name: Natalia Cáceres Arandía Role: annotator</p> <p>Actor Full Name: Natalia Cáceres Arandía</p>	<p>Role: depositor</p> <p>Media Files Filename: Pbh86_Yor05b_Ancha1.w... File handle: http://hdl.handle.net/ File type: Audio Access: U</p> <p>Filename: Pbh86_Yor05b_Ancha1.jp... File handle: http://hdl.handle.net/ File type: Image Access: U</p> <p>Written Resources Filename: Pbh86_Yor05b_Ancha1.e... File handle: http://hdl.handle.net/ File type: ELAN Access: U</p> <p>Filename: Pbh86_Yor05b_Ancha1.p... File handle: http://hdl.handle.net/ File type: Settings Access: U</p>
---	--	---	---

Show more ▼

PARADISEC

paradisec.org.au

The screenshot shows a web browser window displaying the ParadiseC website. The page title is "Story in Sungkung and Salako (Indonesia)". The URL in the address bar is "catalog.paradisec.org.au/collection?id=https://catalog.paradisec.org.au/repository/AA3". The website header includes the ParadiseC logo, a "Home" link, and navigation links for "Search", "Collections", "Items", "Login", and "Help".

The main content area is organized into several sections:

- Description:** A story told in both Sungkung and Salako (Indonesia)
- ID:** https://catalog.paradisec.org.au/repository/AA3
- Identifier:** AA3
- Date Created:** 2008-05-15
- Content Location:** A map of Borneo, Indonesia, with a blue box highlighting the region of Sungkung and Salako. The map includes labels for Kuching, Kota Samarahan, Betong, Serian, Simanggang, and Pontianak. A scale bar shows 100 km and 50 mi. Below the map, it says "This map is not designed or suitable for Native Title research."
- Collector:** [Alexander Adelaar](#)
- In Language:** [Kendayan \(knx\)](#)
- Subject Language:** [Kendayan \(knx\)](#)
- Countries:** Indonesia

On the right side, there are three panels:

- Access:** A red box with a lock icon and the text "Request access or login for this item." Below it is a link "Sign Up or Log In" and the text "As yet unspecified".
- Content:** Shows "Language: Kendayan" and "File Formats: [audio/mpeg](#), [audio/wav](#)".
- Retrieve Metadata:** Contains links "Download metadata" and "Open metadata in a new window".

TLA

archive.mpi.nl

[ARCHIVE](#) / [DOBES ARCHIVE](#) / [BAURE](#) / [3. ETHNOGRAPHIC DATA](#) / [3.1 DAILY LIFE](#)
 / [CHICHA](#) / [GP-090506P](#)

GP-090506P



Details

Title	GP-090506P
Contributor	Franziska Riedel
Country	Bolivia
Genre	Discourse
Format	audio/x-wav text/x-eaf+xml
Language	Baure
Persistent Identifier	https://hdl.handle.net/1839/00-0000-0000-001B-A8E3-3
Description	GP explains the process of making chicha. GP explica como se prepara la chicha.

Downloadable metadata

[GP-090506P DC](#)
[GP-090506P CMD](#)

Detailed Metadata

[expand all](#)

- ▶ [lat-session](#)

Part of: [GP-090506P](#) (2 objects)

Next

[GP-090506P.eaf](#) [Open](#)

[GP-090506P.WAV](#) [Open](#)

Name : GP-090506P

Title :

Date : 2009-05-06

▼ descriptions

Description : GP explains the process of making chicha.

Description : GP explica como se prepara la chicha.

▼ Location

Continent : South-America

Country : Bolivia

Region : Beni

Address : Baures

▶ Project

▶ Content

▼ Actors

▶ Actor

▼ Actor

Role : Consultant

Name : GP

FullName : GP

Code : GP

FamilySocialRole : Unspecified

EthnicGroup :

BirthDate : 1938

Sex : Female

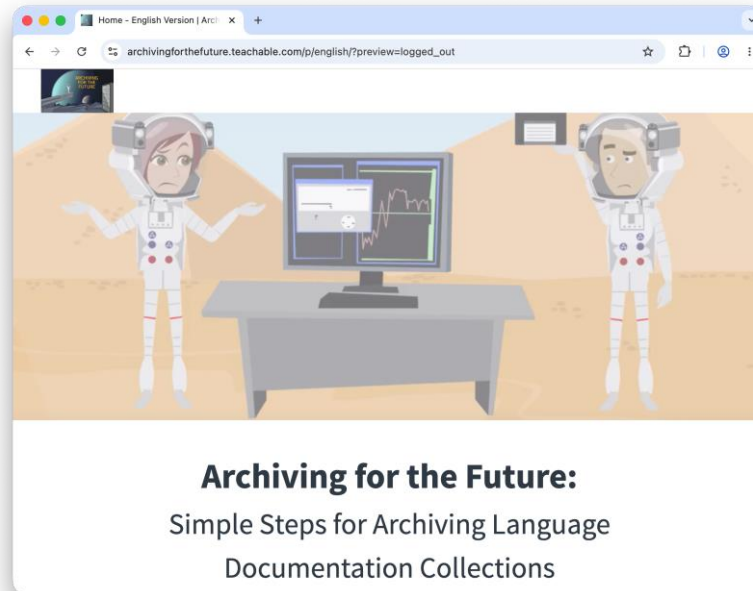
Collecting metadata

- Collecting metadata during data collection saves time in the long run and typically results in better metadata quality
- Some specialized tools exist, e.g. Lameta (<https://sites.google.com/site/metadatatooldiscussion/home>)
- Using a spreadsheet for metadata collection and enter/convert into required format at a later stage is also an option. AI models can be helpful for this work.

Further resources

Archiving for the Future

- Free online course about archiving language documentation collections, created by Susan Smythe Kung et al.:
<https://archivingforthefuture.teachable.com>



Archive-specific resources

- ELAR Resources for Language Documentation and Archiving:
<https://www.elararchive.org/dk0000>
- The Language Archive deposit manual and screencasts:
<https://archive.mpi.nl/tla/deposit-manual-tla>
<https://archive.mpi.nl/forums/c/tla/archiving-info/9>
- PARADISEC deposit information:
<https://www.paradisec.org.au/deposit/>

Questions?



Paul.Trilsbeek@mpi.nl

archive.mpi.nl