MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

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

The Language Archive
CLARIN B CENTRE
Clariah
Funded by the European Union

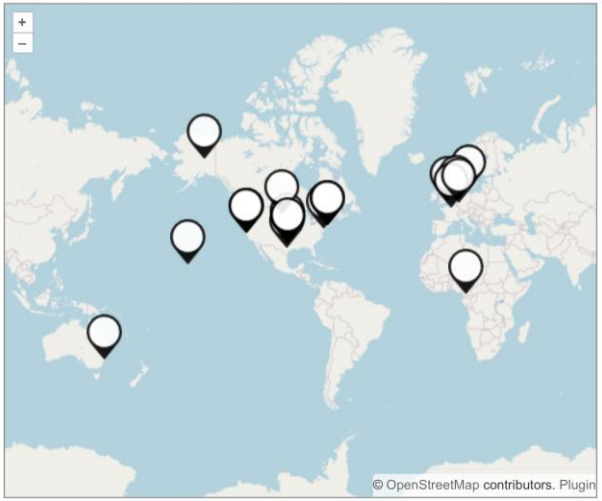
Finding Language Resources

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

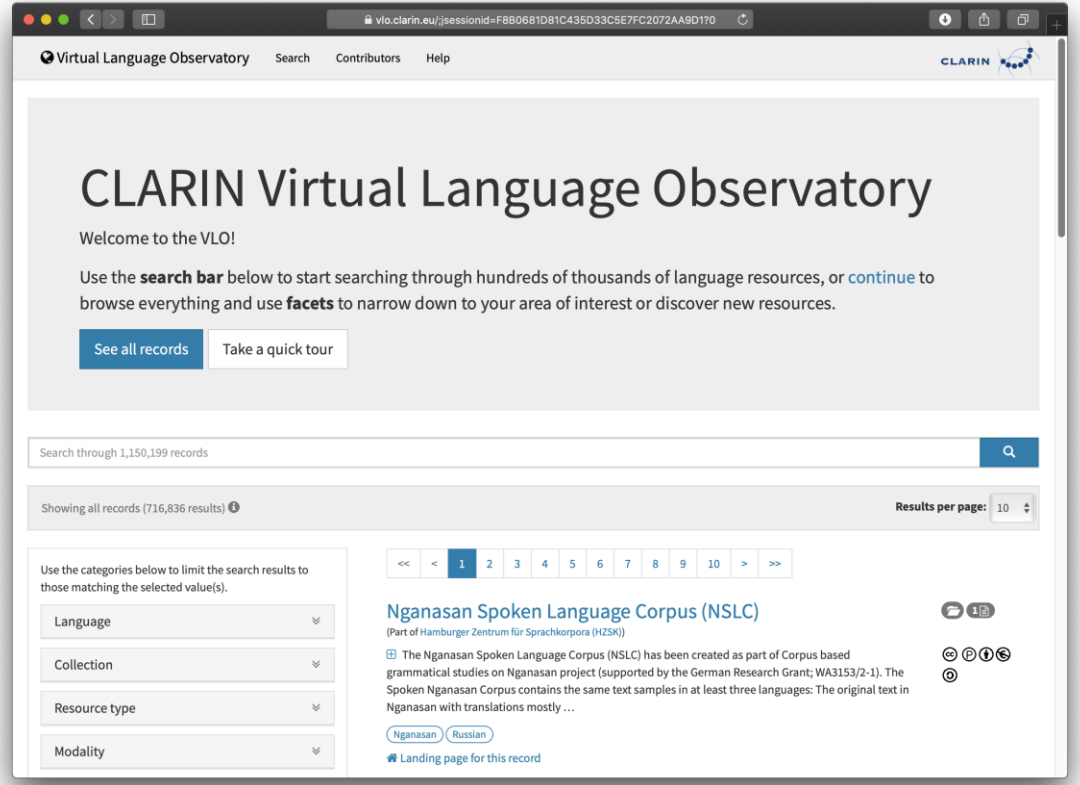
CLARIN Centres

- 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

FOSTERLANG

MAX PLANCK

CLARIN Virtual Language Observatory (VLO)



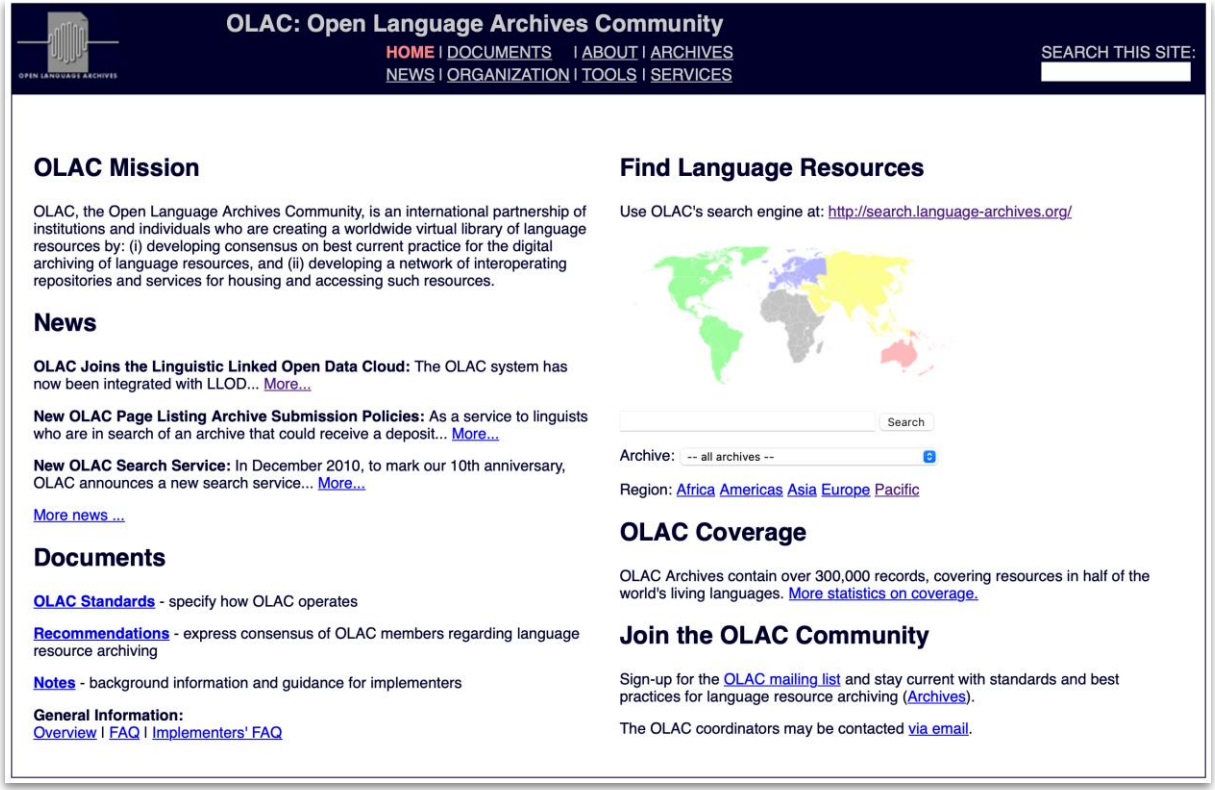
FOSTERLANG

MAX PLANCK

CLARIN Virtual Language Observatory (VLO)

vlo.clarin.eu

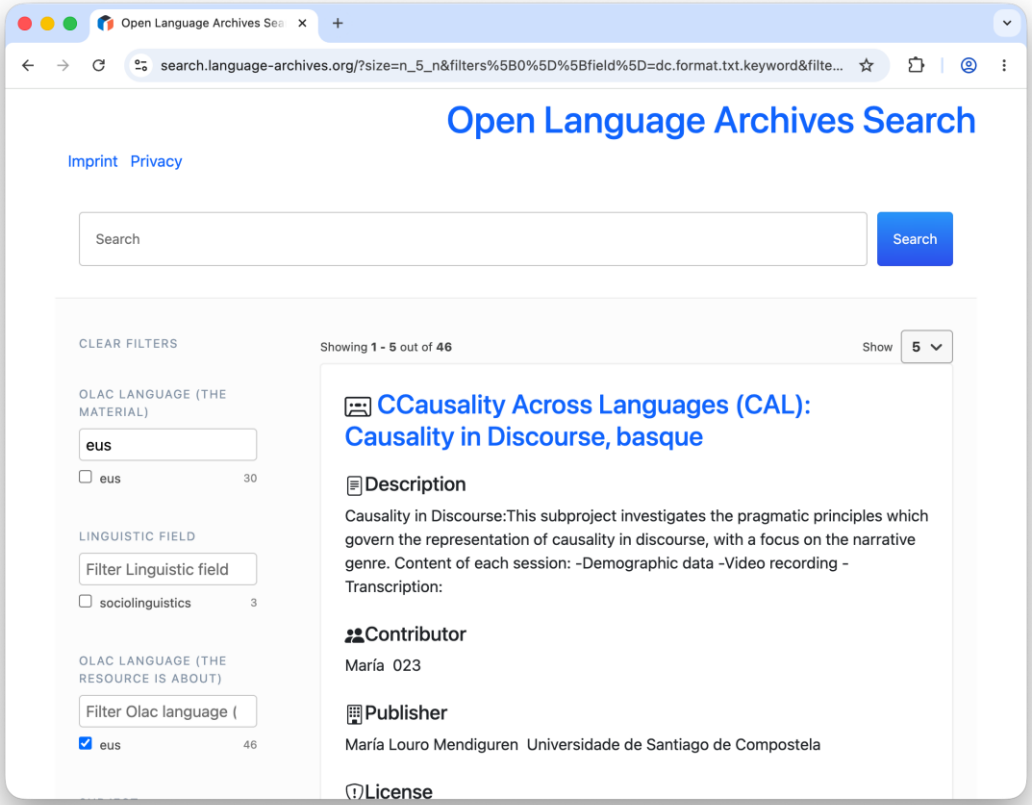
Open Language Archives Community (OLAC)



FOSTERLANG

MAX PLANCK

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



FOSTERLANG

MAX PLANCK

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

Source: wikipedia.org

FOSTERLANG

MAX PLANCK

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

FOSTERLANG

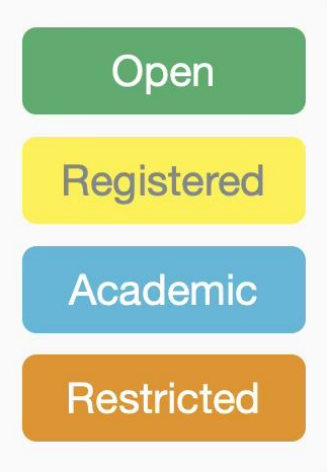
glottolog.org

MAX PLANCK

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



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 an 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

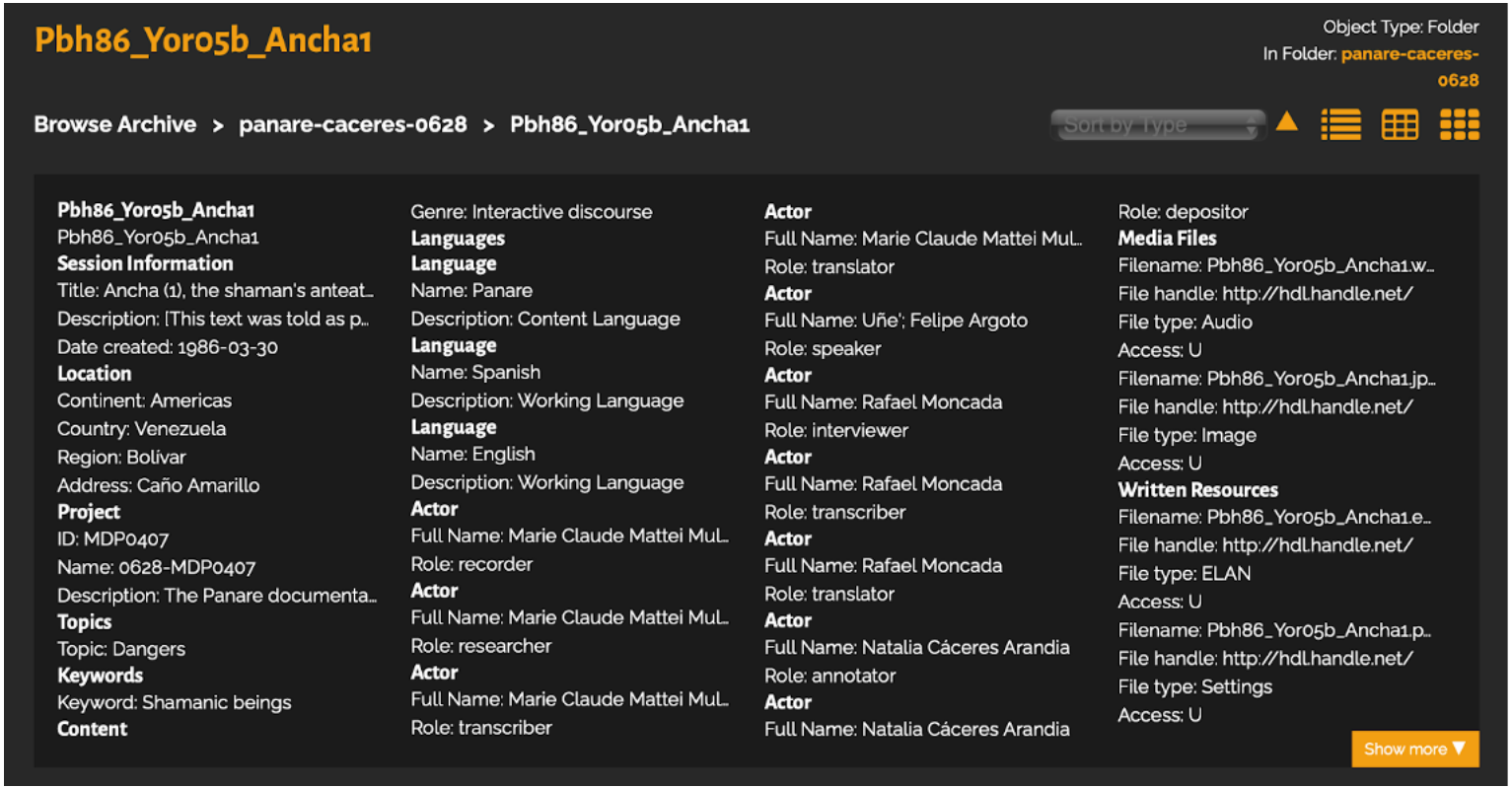
- Component MetaData Infrastructure, developed within CLARIN
- Not a single schema for metadata, but a framework that enables the composition of metadata “profiles” from building blocks that are stored in a central registry
- Enables the creation of metadata profiles that have been tailored for specific 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

ELAR



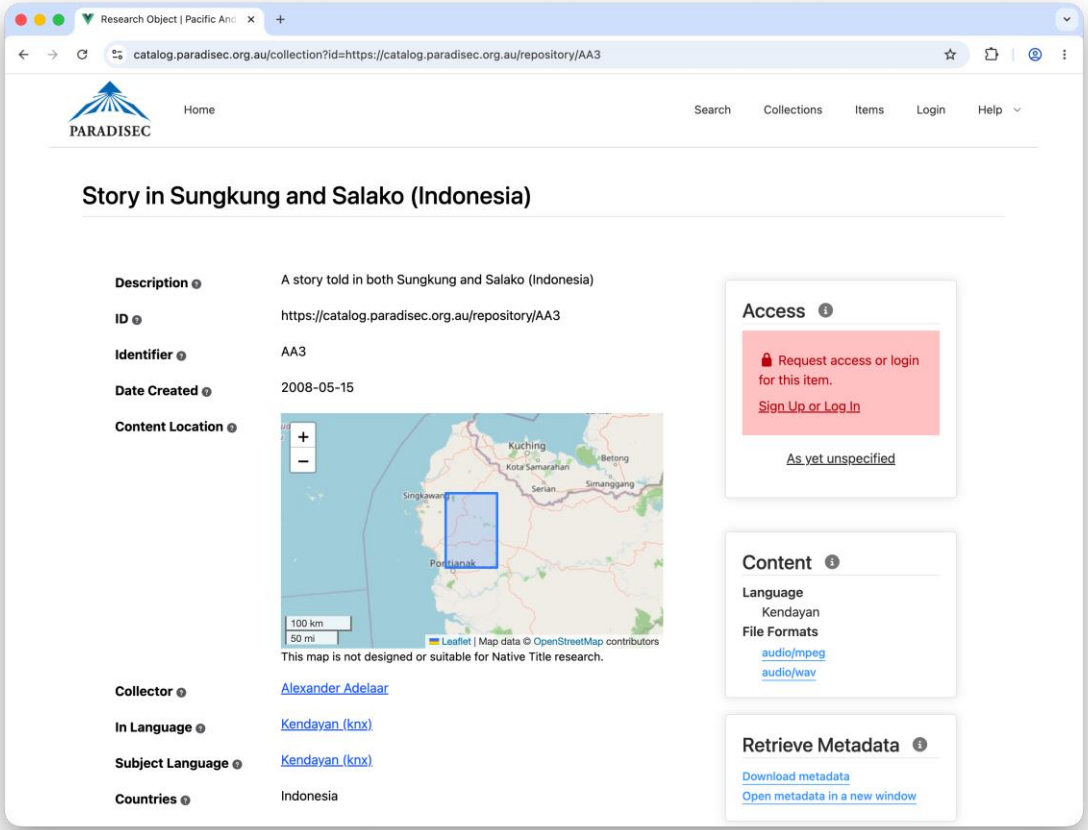
FOSTERLANG

www.elararchive.org

MAX PLANCK

PARADISEC

paradisec.org.au



FOSTERLANG

MAX PLANCK

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

FOSTERLANG

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

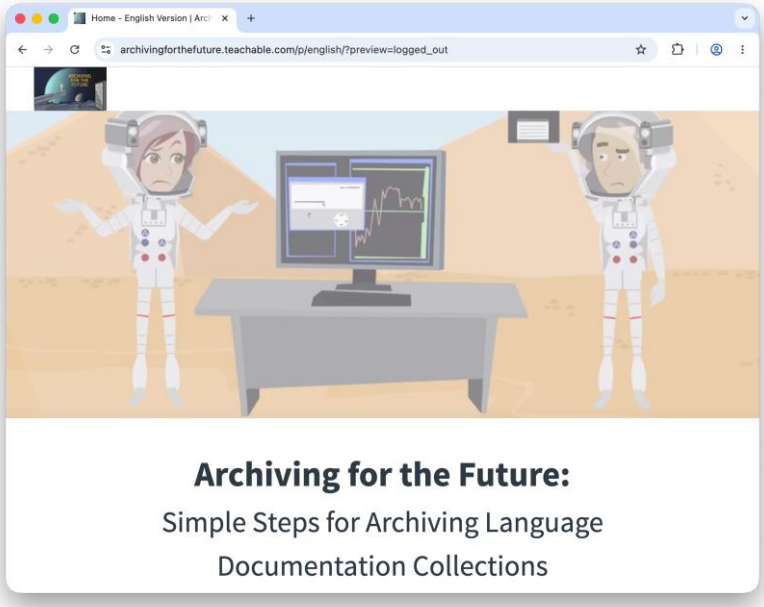
Collecting metadata

- 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?

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

WWW.MPI.NL

Paul.Trilsbeek@mpi.nl

archive.mpi.nl