"Pre-archiving" with the California Language Archive: Incremental archiving and early ongoing curation

Andrew Garrett, Edwin Ko, Julia Nee,
Zachary O'Hagan & Ronald Sprouse

University of California, Berkeley

ICLDC 6, Honolulu
3 March 2019

Acknowledgement

Hanau ka Honu{a} noho i kai,
Kia'i ia e ke Kuhonua noho i uka.
"Born is the turtle living in the sea;
Guarded by the *Maile* seedling living on land."
— *Kumulipo* 414–415 (tr. Beckwith)

Traditional archiving steps

1. Do PhD work.

Traditional archiving steps

1. Do PhD work.
2. Get an academic position.

Traditional archiving steps

1. Do PhD work.
2. Get an academic position.
3. Have a productive research and teaching career.

Traditional archiving steps

1. Do PhD work.
2. Get an academic position.
3. Have a productive research and teaching career.
4. Die after a full life, surrounded by friends and family. They love you and will miss you.

Traditional archiving steps

1. Do PhD work.
2. Get an academic position.
3. Have a productive research and teaching career.
4. Die after a full life, surrounded by friends and family. They love you and will miss you.
5. Also, they give your books, papers, and boxes of recordings to some university archive.

Traditional archiving steps

1. Do PhD work.
2. Get an academic position.
3. Have a productive research and teaching career.
4. Die after a full life, surrounded by friends and family. They love you and will miss you.
5. Also, they give your books, papers, and boxes of recordings to some university archive.
6. Later, 50–75 years after you did the work, the archive gets a grant to catalog your stuff.

Traditional archiving steps

1. Do PhD work.
2. Get an academic position.
3. Have a productive research and teaching career.
4. Die after a full life, surrounded by friends and family. They love you and will miss you.
5. Also, they give your books, papers, and boxes of recordings to some university archive.
6. Later, 50–75 years after you did the work, the archive gets a grant to catalog your stuff.
7. One day a graduate student finds out about your collection, and decides to work with your materials. (Repeat.)

Modern digital archiving: Typical steps

1. Apply for a grant, and get it.

Modern digital archiving: Typical steps

1. Apply for a grant, and get it.
2. Do documentation (and analysis) for a few years.

Modern digital archiving: Typical steps

1. Apply for a grant, and get it.
2. Do documentation (and analysis) for a few years.
3. Put digital files in folders on a computer.

Modern digital archiving: Typical steps

1. Apply for a grant, and get it.
2. Do documentation (and analysis) for a few years.
3. Put digital files in folders on a computer.
4. Back up computer files.

Modern digital archiving: Typical steps

1. Apply for a grant, and get it.
2. Do documentation (and analysis) for a few years.
3. Put digital files in folders on a computer.
4. Back up computer files.
5. At the end of the project, some years after you started, locate and rename your digital files, reorganize them, create all-new metadata documents in required formats, and submit to an archive.

Modern digital archiving: Five problems

- Overwhelmingness

  The quantity of material and amount of work are daunting (archives can be complicated), so people do not archive.

- Duplicated effort

  Organizing research files and an archival collection are partly duplicative activities.

- Lapsus memoriae

  Metadata accuracy may deteriorate over years.

- Access negotiations

  It might be harder to work out restrictions and exclusions after several years.

- Disaster

  While you're waiting: backup failure! fire! flood! theft!

California Language Archive (CLA) contents

The CLA is the online catalog and content portal for the Survey of California and Other Indian Languages, Department of Linguistics, University of California, Berkeley. As of early 2019:

- ca. 350 linear feet of boxes of paper materials and analog sound recordings, from 280–464 languages, depending how you count
- ca. 14,750 accessioned items visible in the catalog
- ca. 33,000 digital files (∼90% are audio files; some content is restricted-access)

Creation of the CLA was supported by the National Endowment for the Humanities (Documenting Endangered Languages, grant PD-50005-07).
Web address: `http://cla.berkeley.edu/`.

Long-term preservation, accessibility, and discoverability

Despite the "string and chewing gum" problem (Thieberger 2019), CLA provides what an archive should:

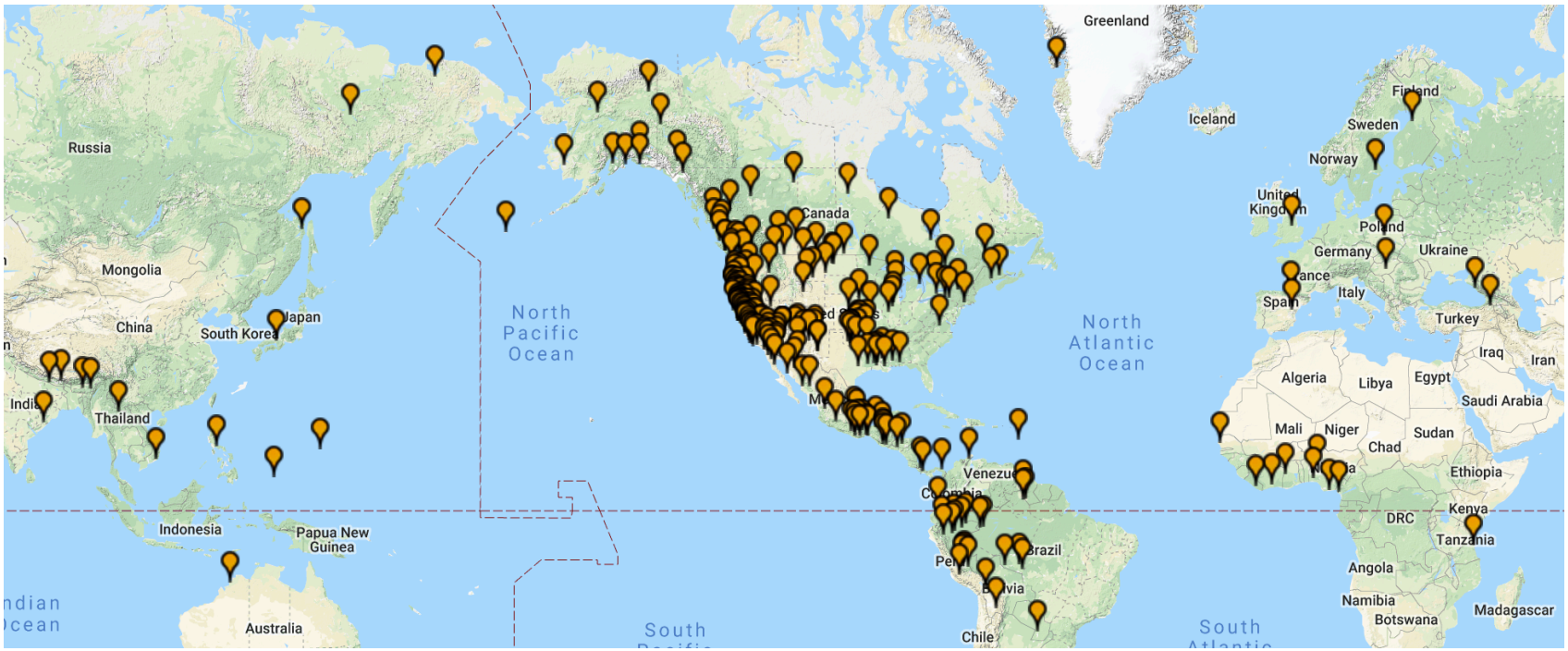
- digital content assigned to items (file bundles) with permanent Digital Object Identifiers
- digital content archived in Merritt, a digital preservation repository maintained by the University of California Curation Center at the California Digital Library (duplicate copies stored in Amazon S3)
- participates in DELAMAN (`http://www.delaman.org/`), the Digital Endangered Languages and Musics Archives Network
- content metadata contributed to OLAC, the Open Language Archives Community (`http://www.language-archives.org/`)

Selected relevant literature: Bird & Simons 2003, Johnson 2004, Robinson 2006, Thieberger 2010, Conathan 2011, Woodbury 2014, Henke & Berez-Kroeker 2016, Berez-Kroeker & Henke 2018

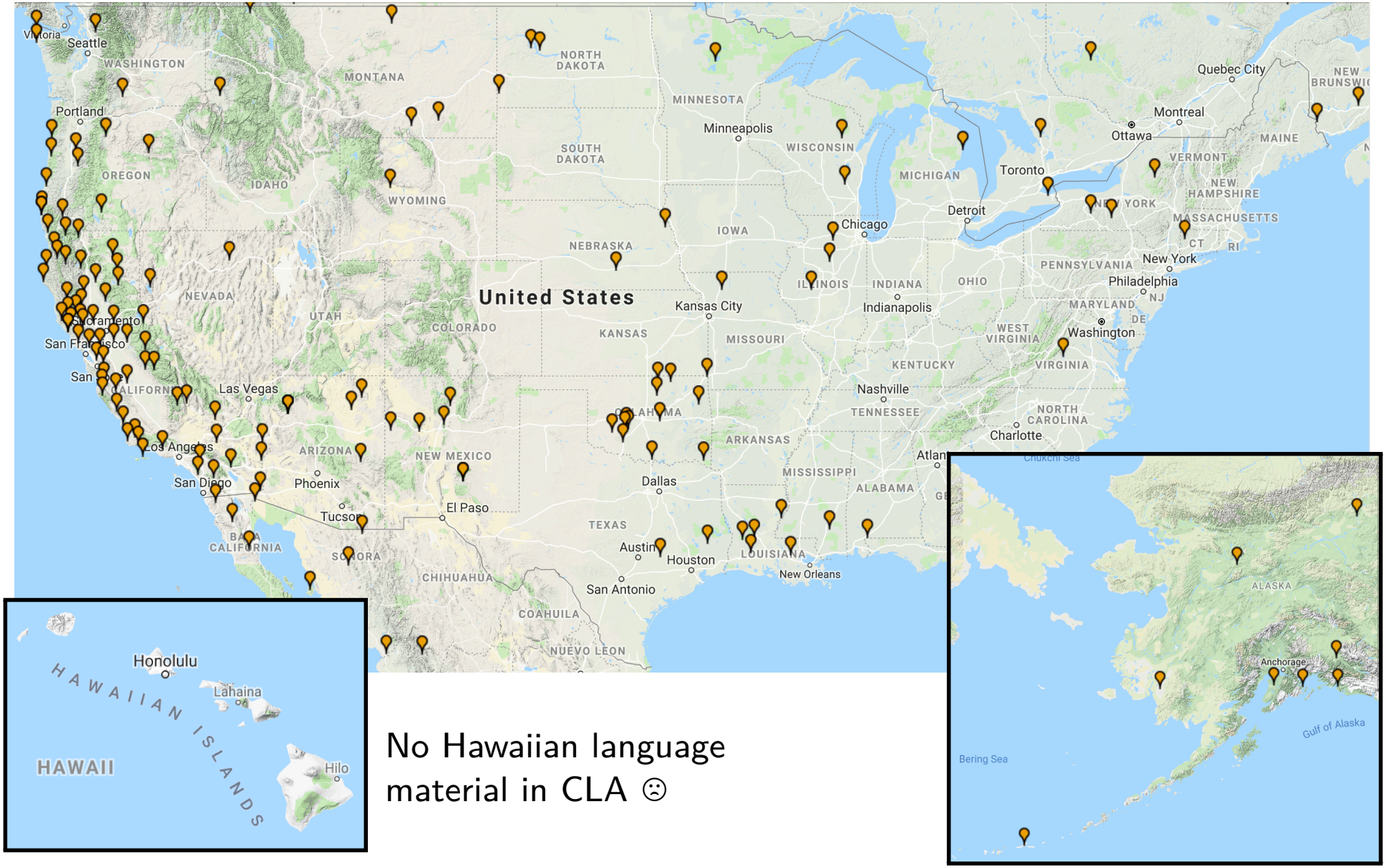
Analog (wax, wire, and tape) recordings & coauthors

CLA map interface: World view

Most CLA items are associated with languages spoken in the western hemisphere (with a historical focus on California and western North America). Each pin is a language:

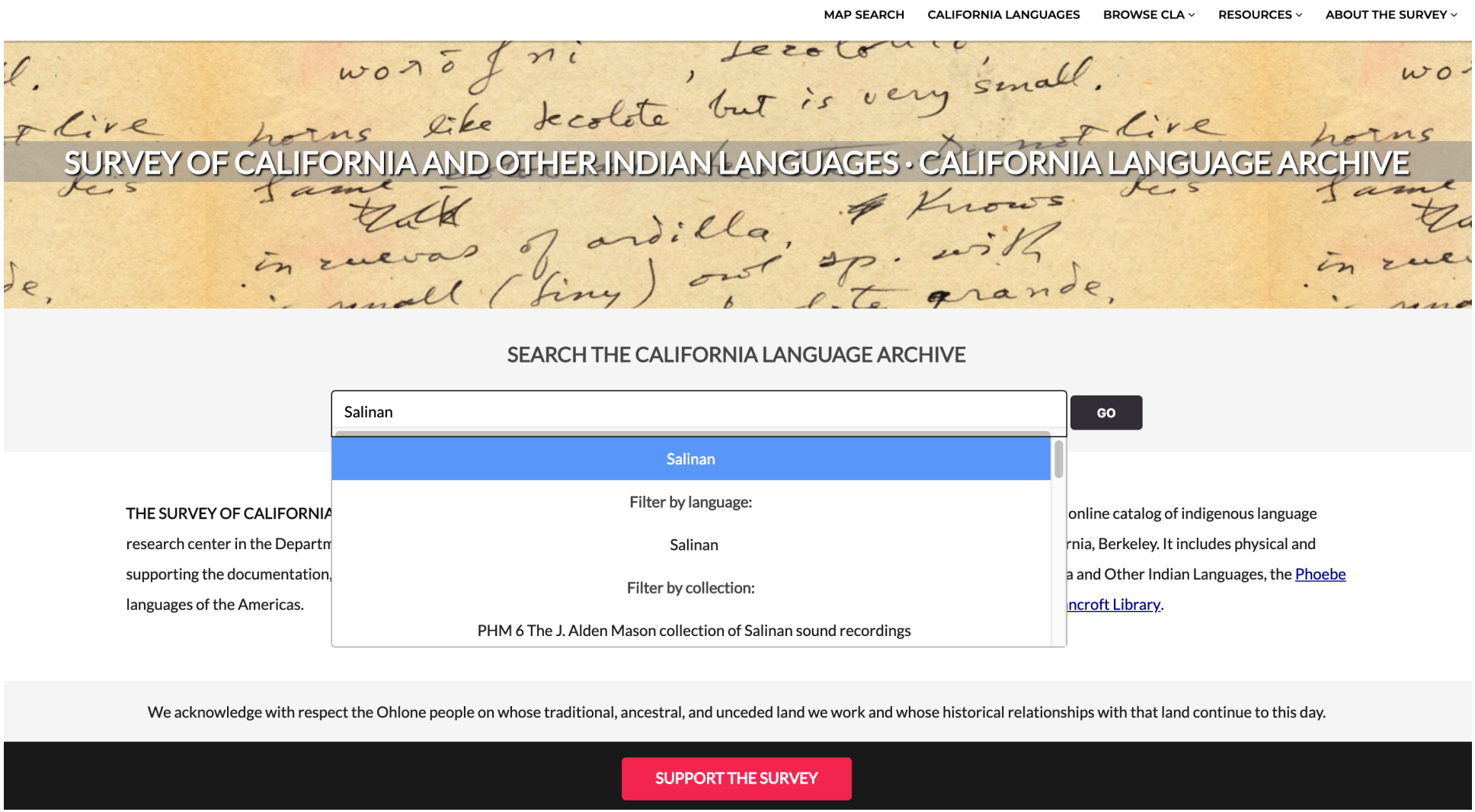


CLA map interface: US view



No Hawaiian language material in CLA ☹

CLA search


MAP SEARCH
CALIFORNIA LANGUAGES
BROWSE CLA
RESOURCES
ABOUT THE SURVEY
SURVEY OF CALIFORNIA AND OTHER INDIAN LANGUAGES · CALIFORNIA LANGUAGE ARCHIVE
SEARCH THE CALIFORNIA LANGUAGE ARCHIVE
Salinan
GO
Salinan
Filter by language:
Salinan
Filter by collection:
PHM 6 The J. Alden Mason collection of Salinan sound recordings
THE SURVEY OF CALIFORNIA
online catalog of indigenous language
research center in the Departm
rnia, Berkeley. It includes physical and
supporting the documentation,
a and Other Indian Languages, the Phoebe
languages of the Americas.
ncroft Library.
We acknowledge with respect the Ohlone people on whose traditional, ancestral, and unceded land we work and whose historical relationships with that land continue to this day.
SUPPORT THE SURVEY

CLA search results

YOU SEARCHED FOR: SALINAN

Collections (7) | Items (32)

1 - 25 of 32 results

Show All/Collapse All

1. Aspects of Salinan grammar (1987)
2. California Indian Languages
3. Conversation in Salinan (13 Sep 1954) (1 digital file, with audio)
4. Discussion of an unidentified topic (08 Sep 1954) (1 digital file, with audio)
5. Elicitation of loanwords (08 Sep 1954) (1 digital file, with audio)
6. Elicitation of minimal pairs (13 Aug 1954) (1 digital file, with audio)
7. Elicitation of numbers 1-30 and 100. Also includes counting adverbials. (30 Jun 1954) (1 digital file, with audio)
8. Elicitation of words and phrases like 'good morning' and 'how are you?' (30 Jun 1954) (1 digital file, with audio)
9. English-Inezeno Chumash vocabulary; English-Chumash vocabulary; Twenty-nine words and phrases of the dialect of the San Miguel (Salinan) Indians and four phrases of the San Luis Obispo Obispeno Chumash), given by Rafael Solaris, compared with Tsa-ma-la (Barbareno Chumash) ([undated])
10. Fighting Forest Fires (a narrative in Salinan) (1910 December) (3 digital files, with audio)
11. Miscellaneous words and phrases such as thing, house, and man. Includes pronouns, nature terms, and some questions. (30 Jun 1954) (1 digital file, with audio)
12. My trip to San Francisco (a narrative in Salinan) (1910 December) (3 digital files, with audio)
13. Picture texts (05 Aug 1954) (1 digital file, with audio)
14. Pre-Salinan (26 May 1982) (1 digital file)
15. Responses to stickman figures 1-50 (Jul 1954) (1 digital file, with audio)
16. Salinan Photos (07 Sep 1984) (5 digital files)

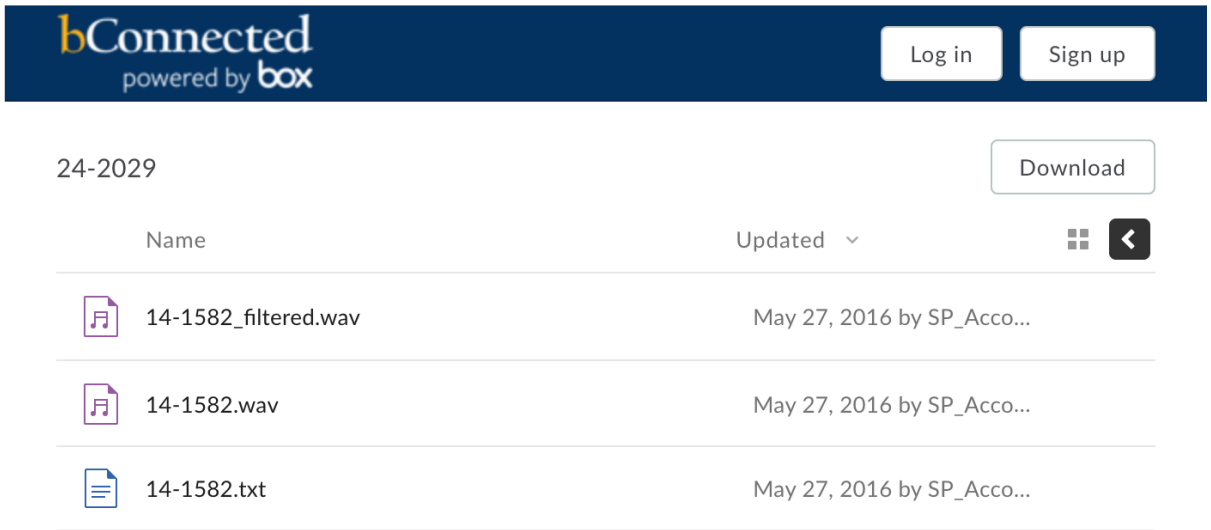
CLA collections & items

MY TRIP TO SAN FRANCISCO (A NARRATIVE IN SALINAN)

Item number: 24-2029
Date: 1910 December
Contributors: Pedro Encinales (consultant), J. Alden Mason (researcher)
Language: Salinan (sln)
Availability: Online access
Description: Keeling catalog note: "Interlinear translation in Mason (1918:98)." Original cylinder 14-1582. 180 speed.
Collection: The J. Alden Mason collection of Salinan sound recordings
Repository: Phoebe A. Hearst Museum of Anthropology
Preferred citation: Pedro Encinales and J. Alden Mason. My trip to San Francisco (a narrative in Salinan), 24-2029, Phoebe A. Hearst Museum of Anthropology, University of California, Berkeley, http://cla.berkeley.edu/item/11145

By using digital assets, you accept our Terms and Conditions.

If files do not appear below, you may also go directly to the asset folder.



Access restrictions

Three access levels for digital content

- **open access**
- **on request** (with documentation of non-commercial purpose)
- **restricted** (depositor permission required)

For example:

ELICITED WORDS AND PHRASES INCLUDING COLORS AND BODY PARTS

Item number: LA1.001
Contributors: Ted Couro (consultant), Margaret Langdon (researcher)
Language: Diegueño (dih)
Availability: Online access by request.
Catalog history: Digital asset LA1.001.001.wav was formerly segment number 004_1.
Place: Campo Indian Reservation
Description: English glosses.
Collection: The Margaret Langdon collection of Diegueño sound recordings
Repository: Survey of California and Other Indian Languages
Preferred citation: Ted Couro and Margaret Langdon. Elicited words and phrases including colors and body parts, LA 1.001, Berkeley Language Center, University of California, Berkeley, http://cla.berkeley.edu/item/17735

Digital assets in this Item (not available for download):

LA1.001.001.wav (74797176 bytes)

Institutional constraints

The creation of a "prearchive" system was partly driven by local circumstances:

- Minimal budget

  We have part of the time of an IT Specialist, and ∼30 hours / week of student researcher help, so cannot have a system that requires extensive staff engagement with each collection.

- Community familiarity

  Most contributors are or were at Berkeley, are close at hand when needed, and form a mutual support network.

  (Many Berkeley linguistics graduate students do fieldwork. Our prearchive is part of a required graduate field methods course.)

We wanted something easy for depositors and for us.

How to begin

Three easy steps:

1. Communicate with the Survey of California and Other Indian Languages. Tell us what kinds of materials you (will) have.

How to begin

Three easy steps:

1. Communicate with the Survey of California and Other Indian Languages. Tell us what kinds of materials you (will) have.
2. Fill out an online form identifying languages, people, and places.

How to begin

Three easy steps:

1. Communicate with the Survey of California and Other Indian Languages. Tell us what kinds of materials you (will) have.
2. Fill out an online form identifying languages, people, and places.
3. We authorize you to use the pre-archive. (Multiple people can be authorized for a single collection.)

How to begin

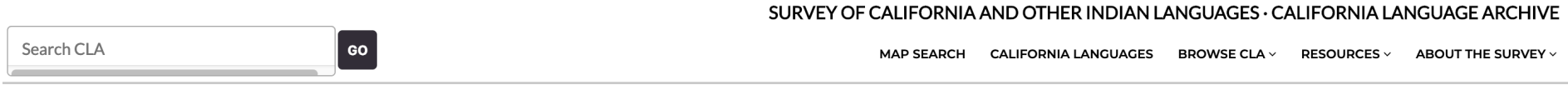
Three easy steps:

1. Communicate with the Survey of California and Other Indian Languages. Tell us what kinds of materials you (will) have.
2. Fill out an online form identifying languages, people, and places.
3. We authorize you to use the pre-archive. (Multiple people can be authorized for a single collection.)

Everything is online — no spreadsheets! Two notes:

- If you're applying for a grant, please include funding to help with our costs ($500 per collection + $1k per terabyte).
- We do archive non-digital materials, obviously not through a "prearchive." Let us know if you're depositing notebooks, etc.

Accessing your predepositor page



DEPOSITOR LOGIN

FOR DEPOSITORS: DONATING MATERIAL TO THE BERKELEY LANGUAGE ARCHIVES

The Survey of California and Other Indian Languages supports the documentation, study, and revitalization of the indigenous languages of California and the Americas. We maintain a major archive of field notes and other documentary materials, some of it digitized and available online; we also curate the collection of linguistic field recordings in the Berkeley Language Center. For more information about the scope and nature of our collections, please read our Collection and Mission webpages.

Donations of materials within the scope of our collections policy are welcome. If you have material you would like to donate to the Survey of California and Other Indian Languages, or to the Berkeley Language Center, please contact the Survey Director, Andrew Garrett, by telephone, email, or letter.

In an effort to defray the costs that are involved in maintaining the collections in our archive, we kindly ask that depositors, espcially those with funding from grants agencies, consider donating to the Survey:

- $500 per collection,
- plus $1000 per terabyte,
- plus paid Box account if files greater than 250MB per file.

Your collections

CLA PREARCHIVE [**Help guide**]

Collections for Julia Eileen Nee

Click on a Collection identifier to edit metadata for the Collection and to find file bundles contained in the Collection.

1. Identifier: SCL 2016-02
   Title: Teotitlán del Valle Zapotec
2. Identifier: SCL 2016-13
   Title: Berkeley Field Methods: South Bolivian Quechua
3. Identifier: SCL 2017-02
   Title: TdVZ Grammar Sketch
4. Identifier: SCL 2017-03
   Title: Path in South Bolivian Quechua of Cochabamba

Editing collection-level metadata

CLA PREARCHIVE [Help guide]

Collection SCL 2016-02

Create a new file bundle | Set up Contributions, Languages, or Places for this Collection

Title

Teotitlán del Valle Zapotec

Associated materials

Historical information

Teotitlán del Valle Zapotec (ZAPZ) is an endangered language spoken in Teotitlán del Valle, approximately 25 kilometers east of Oaxaca City, the capital of Oaxaca state in southern Mexico. Government estimates as of 2010 stated that there were 198 monolingual speakers and 3,601 bilingual speakers. Although the language is categorised by Ethnologue as being 'developing' because the community is in the process of implementing an orthography and of creating a literature, one can observe that the youngest generation is not becoming fluent in the language. While there is bilingual education available in preschool and high school, these programs do not effectively teach the language to students. Instead, most children become fluent in Spanish, the regionally dominant language and the language used in the elementary school. The materials in this collection were developed primarily by Julia Nee as part of ongoing fieldwork

Scope and content

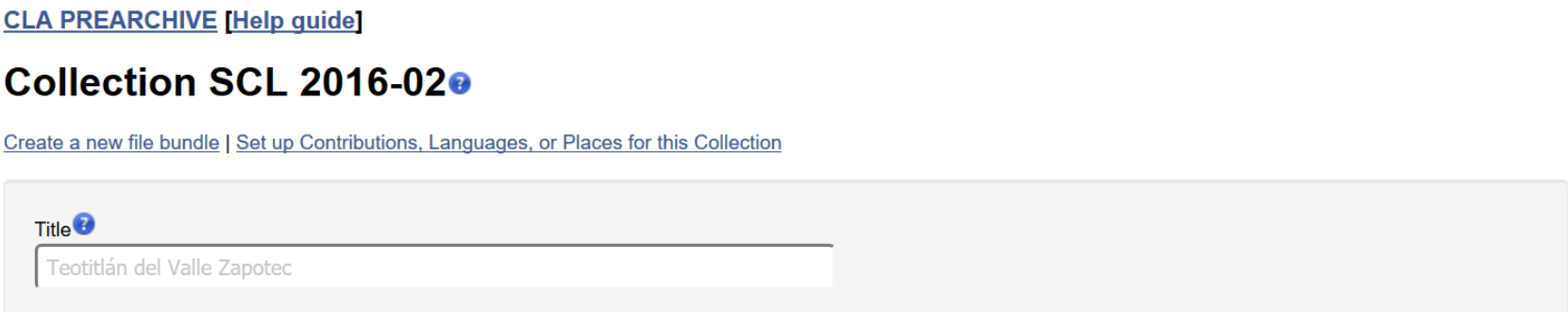
Audio recordings of elicitation sessions and of conversational texts; field notes; ancillary documents

Update descriptive metadata

Beyond collection-level content metadata

Congratulations on your awesome new collection! What's next?

- Add items (file bundles) to the collection?
- Associate names (of languages, people, and places) with the collection?



Collection-level contributions

CLA PREARCHIVE [Help guide]

Set up Contributions, Languages, and Places for SCL 2016-02

Names on these lists will be available when editing your Collection and File Bundles. Add names that you want to be available for associating with your Bundles.

Contribution list

- Bazán Chávez, Enedina (consultant)
- Bazán Chávez, Verónica (consultant)
- Bazán Chávez, Janet (consultant)
- Chávez, Tomasa (consultant)
- Chávez, Juana (consultant)
- Lazo Martínez, Manuel (consultant)
- Lazo Martínez, Constantino (consultant)
- Lazo Martínez, Isabel (consultant)
- Lazo Pérez, Efraín (consultant)
- López Montaño, Teresa (consultant)
- Martinez, Sergio (consultant)
- Martinez Sosa, Trinidad (consultant)
- Martínez, Pedro (consultant)
- Nee, Julia Eileen (researcher)

Person

Role: ---- Choose a role ---

Add contribution

Language list

- Teotitlán del Valle Zapotec

Language

Add language

Place list

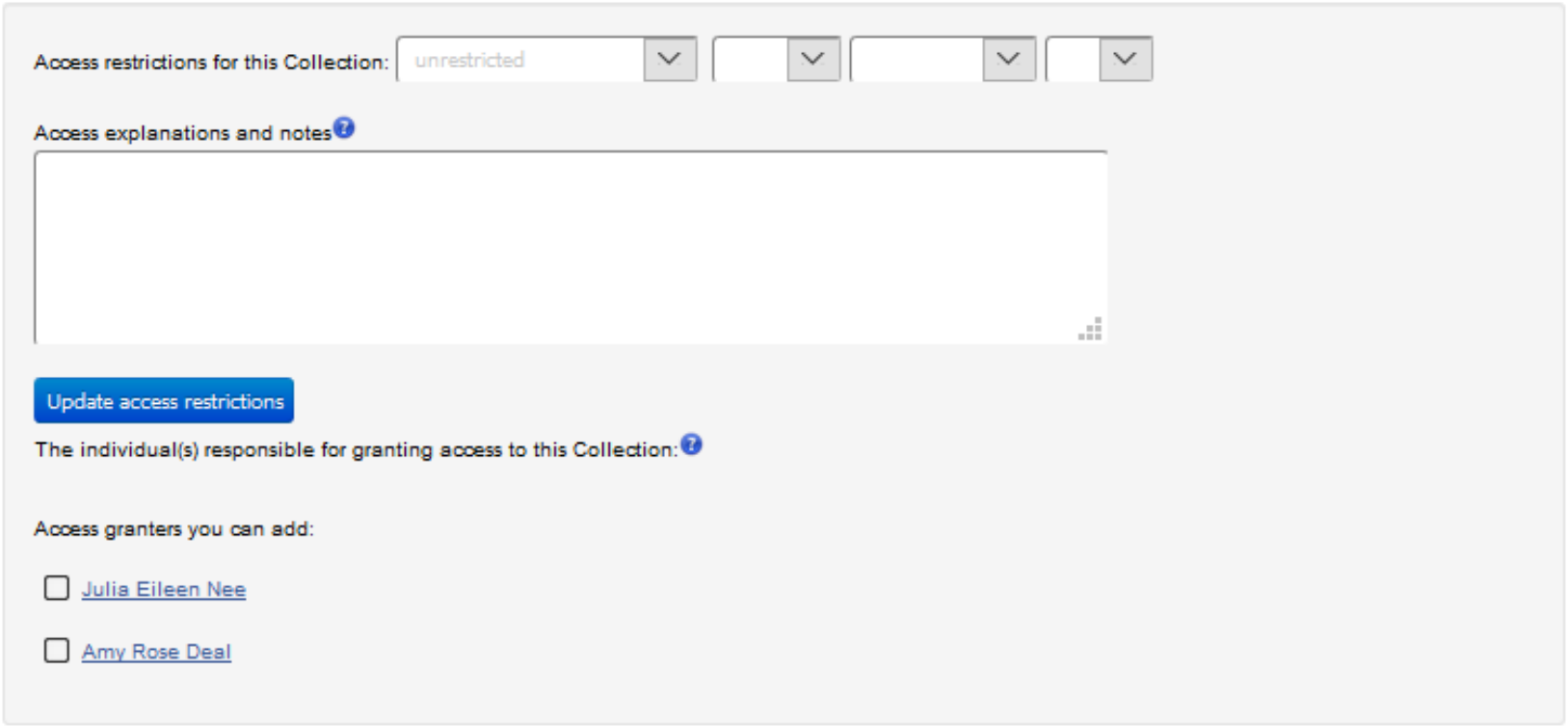
- Teotitlán del Valle, Oaxaca, Mexico

Place

Add place

(If autocomplete boxes aren't working, try refreshing the page.)

Collection-level access restrictions



Item-level (file bundle) metadata: Content

CLA PREARCHIVE [Help guide]

Metadata for file bundle 2016-02.001

Manage files in this bundle | Box report | Return to collection SCL 2016-02

Short Title

Definiteness elicitation with three consultants for Teotitlan del Valle Zapotec

Date(s)

YYYYMMDD-YYYYMMDD | 2016 | January | 5 – 2016 | January | 15

Description

Part of project on variation in definiteness cross-linguistically.

OLAC descriptors:

Primary Text: ◉ No ○ Yes | Lexicon: ◉ No ○ Yes | Language Description: ○ No ◉ Yes | Contains texts: ◉ No ○ Yes

Update descriptive metadata

Item-level (file bundle) metadata: Access; contributions

Access restrictions for this File Bundle: are inherited from the Collection

Update access restrictions

The individual(s) responsible for granting access to this File Bundle: are inherited from the Collection

Contributions:

- [x] Julia Eileen Nee (researcher)
- [x] Enedina Bazán Chávez (consultant)
- [x] Sergio Martinez (consultant)
- [x] Tomasa Chávez (consultant)

Contributions you can add:

- [ ] Verónica Bazán Chávez (consultant)
- [ ] Janet Bazán Chávez (consultant)
- [ ] Juana Chávez (consultant)
- [ ] Manuel Lazo (consultant)

Languages:

- [x] Teotitlán del Valle Zapotec

Places:

- [x] Teotitlán del Valle, Oaxaca, Mexico

Relations:

This file bundle:
No relations are associated with this file bundle.

Relations you can add:

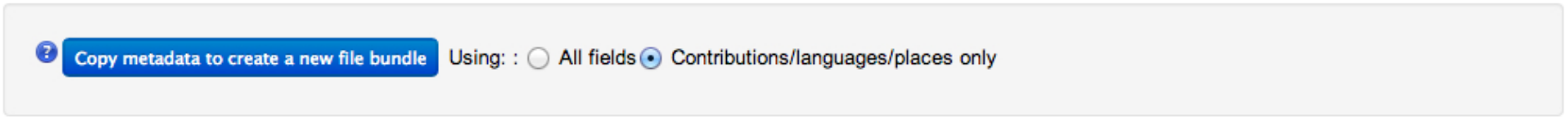
This bundle: ---- Choose a relation ----

- [ ] 2016-02.004 Field Notes Notebook 1 (Jan 2016 to Jul 2016)
- [ ] 2016-02.005 Questions about local and non-local topics (30 May 2016)
- [ ] 2016-02.006 Transcription of the text in track 2016-05-30-Pedro-Martinez-1.wav (31 May 2016)
- [ ] 2016-02.007 Elicitation of simple sentences (31 May 2016)

Cloning items (file bundles) to create new items

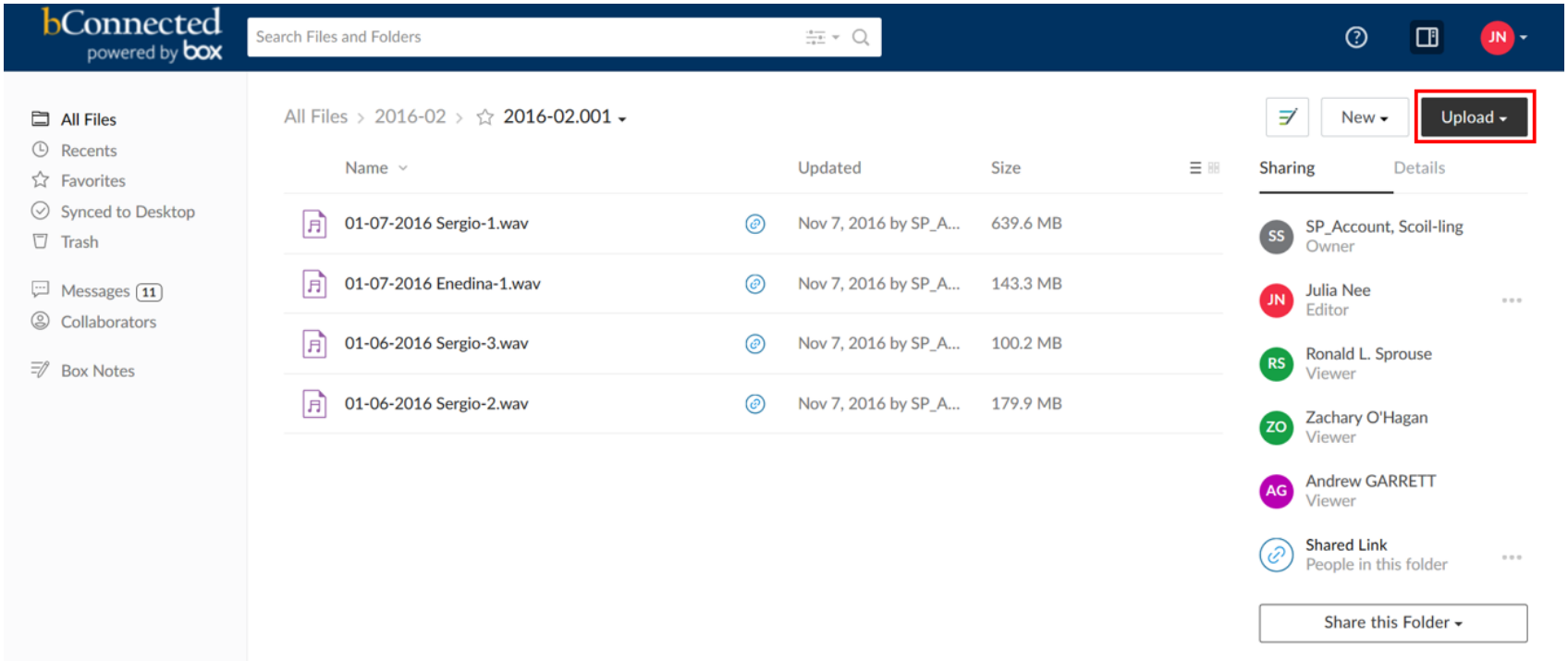
Do you have two or more similar bundles to create?

- Same language(s)?
- Same people?
- Same place(s)?



Uploading digital files to items (file bundles)

You can drag and drop, or use the upload button (highlighted in red).



How to accession

Three easy steps:

1. Communicate with the Survey of California and Other Indian Languages.

How to accession

Three easy steps:

1. Communicate with the Survey of California and Other Indian Languages.
2. Sign a gift agreement; transfer funds from your grant, etc.

How to accession

Three easy steps:

1. Communicate with the Survey of California and Other Indian Languages.
2. Sign a gift agreement; transfer funds from your grant, etc.
3. We push a button. Your catalog information (and digital content, as appropriate) becomes visible to the world. Mazel tov!

How to accession

Three easy steps:

1. Communicate with the Survey of California and Other Indian Languages.
2. Sign a gift agreement; transfer funds from your grant, etc.
3. We push a button. Your catalog information (and digital content, as appropriate) becomes visible to the world. Mazel tov!

Or change your mind; no problem.*

*"Sorry, CLA dudes, your prearchive is awesome but I decided to archive elsewhere after all. I just really love spreadsheets!"

Usage data

CLA material accessioned via the prearchive (2015–2019)

- 27 collections
- 893 items (file bundles)
- ∼11,030 digital files

CLA material currently in the prearchive (to be accessioned)

- 24 collections (some are still empty "placeholders")
- ∼1,100 digital files

Possible concerns

- Reliance on Box to serve files to users; eventually Box will expire, and we'll have to develop a new pathway.
  - Archival durability unaffected; permanent preservation copies maintained in the UC Merritt repository.
- Metadata flexibility may not suit every archive; highly convenient for depositors but inconsistencies may result.
- Works well with a linked community of (often local) depositors; untried with numerous isolated depositors.

Features that have worked for us

- Single intuitive web-based interface
- Continuous incremental uploading of digital files
- Incremental accessioning into the public archive & catalog
- Ongoing curation during the prearchive phase
  - Contributors can change collection metadata, item metadata, and item (file bundle) content.
  - We don't keep track of contributors' changes and they don't have to get approval.

We encourage other archives to adopt aspects of our approach that might work well for them. (We also welcome inquiries from potential depositors: `scoil-ling@berkeley.edu`.)

Mahalo!

References

- Berez-Kroeker, Andrea L., & Ryan E. Henke. 2018. Language archiving. *Oxford handbook of endangered languages*, ed. Kenneth Rehg & Lyle Campbell, pp. 347–369.
- Bird, Steven, & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79: 557–582.
- Conathan, Lisa. 2011. Archiving and language documentation. *Cambridge handbook of endangered languages*, ed. Peter K. Austin & Julia Sallabank, pp. 235–254.
- Henke, Ryan E., & Andrea L. Berez-Kroeker. 2016. A brief history of archiving in language documentation. *LDC* 10: 411–457.
- Johnson, Heidi. 2004. Language documentation and archiving. *LDD* 2: 140–153.
- Robinson, Laura. Archiving directly from the field. *Sustainable data from digital fieldwork*, ed. Linda Barwick & Nicholas Thieberger, pp. 23–32.
- Thieberger, Nicholas. 2010. Anxious respect for linguistic data. *Endangered languages of Austronesia*, ed. Margaret Florey, pp. 141–158.
- Thieberger, Nicholas. 2019. The ongoing challenge of connecting speakers to archival language records. ICLDC 6, 1 March.
- Woodbury, Anthony C. 2014. Archives and audiences. *LDD* 12: 19–36.