Kaipuleohone, Kani‘āina, RDA LDIG & The Austin Principles of Data Citation

Andrea Berez-Kroeker
U Hawai‘i at Mānoa
andrea.berez@hawaii.edu

COCOSDA workshop @ LDC February 9 2018

Kaipuleohone

University of Hawai‘i Digital Language Archive

(‘the gourd of sweet voices’)

UH Digital Language Archive

Administered by Department of Linguistics

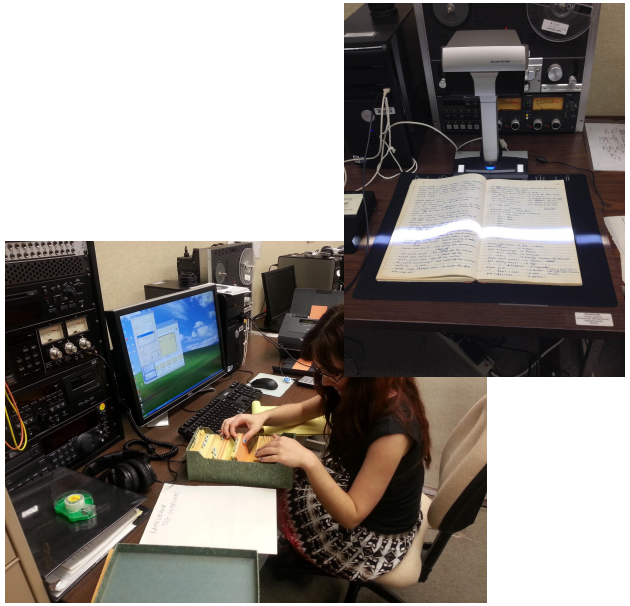
Housed in ScholarSpace, UH IR (DSpace) at Hamilton Library

Started in 2007 by predecessor Nick Thieberger, modeled after PARADISEC

Started to house digitized older collections from faculty

Now mainly pedagogical function - a “teaching archive” for our graduate students

Kaipuleohone
University of Hawai‘i Digital Language Archive



Specs:

53 collections

4,392 records (=bundle of files), 176 languages

OLAC member ⋆⋆⋆⋆⋆, DELAMAN member

~3.5 TB of Audio, video, transcripts, photos, notebooks, some older printed material

How we obtain data: Mostly from our students & faculty, others working in Pacific

How data is accessed: For non-open collections, through permission from depositor … but infrequent :(

Kaipuleohone
University of Hawai'i Digital Language Archive

Part of graduate student training

- Integrated into the 4-semester langdoc/field methods curriculum
  - Students learn data and metadata collection, etc.
- Houses students' summer fieldwork products
- PhD requirement to archive data from dissertations as condition of graduation

Kaipuleohone
University of Hawai'i Digital Language Archive

Contents **are evolving** in a few ways:

Right now mostly unannotated media

Students only recently starting to conscientiously add annotations

Since summer 2017, all new collections have been voluntarily open

Previously nearly all closed for at least 5 years

We updated our embargo policy

...but the students beat us to the punch!

Kaipuleohone
University of Hawai‘i Digital Language Archive

Relation to “the Americas”

Housed at a US university

Holds a few collection of items from the Americas (sign languages of Peru, Hawaiian Creole English, Hawai‘i Sign Language)

Hawaiian is not a language of the Americas (linguistically)
but Hawaiian language education policies under US authority (DoE / ANA)

Similar $$ influence in other “US interest” locations in the Pacific
(CNMI, Guam, FSM, Palau, American Sāmoa)
…for which we have language material

Challenges we face

Funding

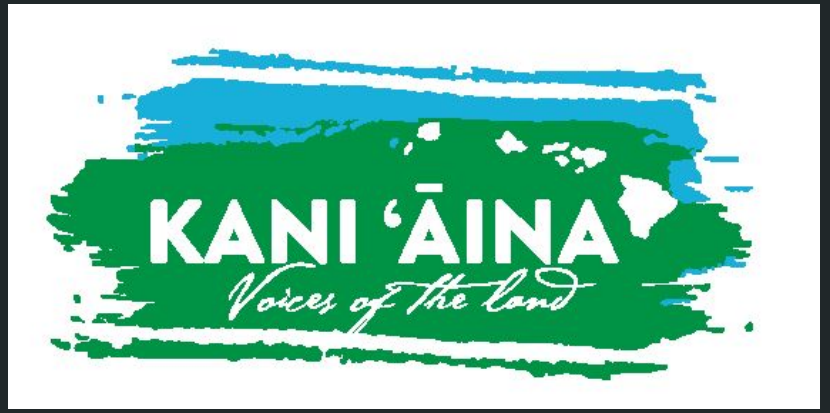
User interface isn't great for searching or showcasing

Current lack of annotation poses problems for repurposing data

www.kaipuleohone.org

Kani‘āina

(‘the reverberating sound of the land’)



National Science Foundation (BCS-1664070) and National Endowment for the Humanities (PD-255910)
PIs Keiki Kawai‘ae‘a, Larry Kimura, Andrea Berez-Kroeker

Kaniʻāina

Joint project between Kaipuleohone and Ka Haka ʻUla O Keʻelikōlani School of Hawaiian Language at UH Hilo (on Big Island)

In-progress (2017-2020)

Repository and access point for 900-1200 hours of transcribed spoken Hawaiian recordings

Builds on celebrated Niupepa collection -

This is the first major collection of spoken Hawaiian

Kaniʻāina

User interface is part of Ulukau, the Hawaiian language digital library (1.4M hits/mo!)

Files preserved in Kaipuleohone

First 40+ hours transcribed in house

- The rest through crowd-sourcing
- Some marked as *kapu* for transcription (reserved for educational purposes)

No English translation provided

Will include outreach to classroom teachers

Kaniʻāina

How we obtain data:

In-house
Plus tight local network
Personal collections, nonprofit collections, governmentally-funded projects…

How data is accessed:

Most will be open (downloadable through Kaipuleohone)
Crucially, with expectation to respect proper *kuleana*

Kaniʻāina

First collection: *Ka Leo Hawaiʻi* radio program

1982-2000, Hosts Larry Kimura (1982-1991) & Puakea Nogelmeier (1991-2000)

Showcases Native Hawaiian conversation (as opposed to neo-Hawaiian)

Alpha version in place http://ulukau.org/kaniaina/cgi-bin/kaniaina

RDA LDIG
&
The Austin Principles of Data Citation in Linguistics

Research Data Alliance Linguistics Data Interest Grp

Extension of a previous NSF project to assess the current state of reproducible research in linguistics (see our position paper)

Aims:

- To foster a culture of reproducible research across linguistics
- To educate colleagues about proper handing / storing / sharing /citing of linguistic data
- To increase the valuation of “data work” in hiring, T&P

Open to anyone: https://www.rd-alliance.org/groups/linguistics-data-ig

Research Data Alliance Linguistics Data Interest Grp

First output:

Austin Principles of Data Citation in Linguistics

The FORCE11 Principles of Data Citation, annotated for linguists

To help linguists understand *why* to cite data

See www.linguisticdatacitation.org to endorse, read FAQs, etc.

Research Data Alliance Linguistics Data Interest Grp

Next intended output:

Gathering / refining / promoting recommended **citation conventions**
for a range of linguistic data types
and tips for using them

First working group meeting planned for March 2018 RDA meeting (Berlin)