Thoughts on language technology and language archiving: South Asia from an outsider perspective

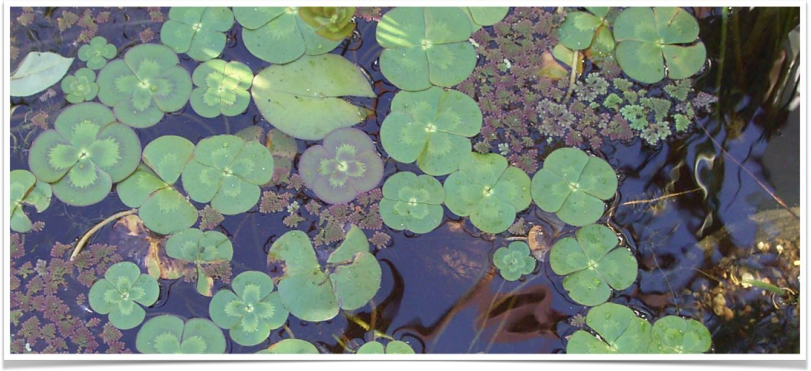
CLAIRE BOWERN
Drawing on work by Sarah Babinski, Irene Yi,
+ KASSANDRA HAAKMAN, JEREMIAH JEWELL, JUHYAE KIM, AMELIA LAKE
YALE UNIVERSITY, DEPT OF LINGUISTICS

ROADMAP

Language archives

- Curation
- Preservation
- Discoverability
- Dissemination

BACKGROUND :: LANGUAGE ARCHIVES



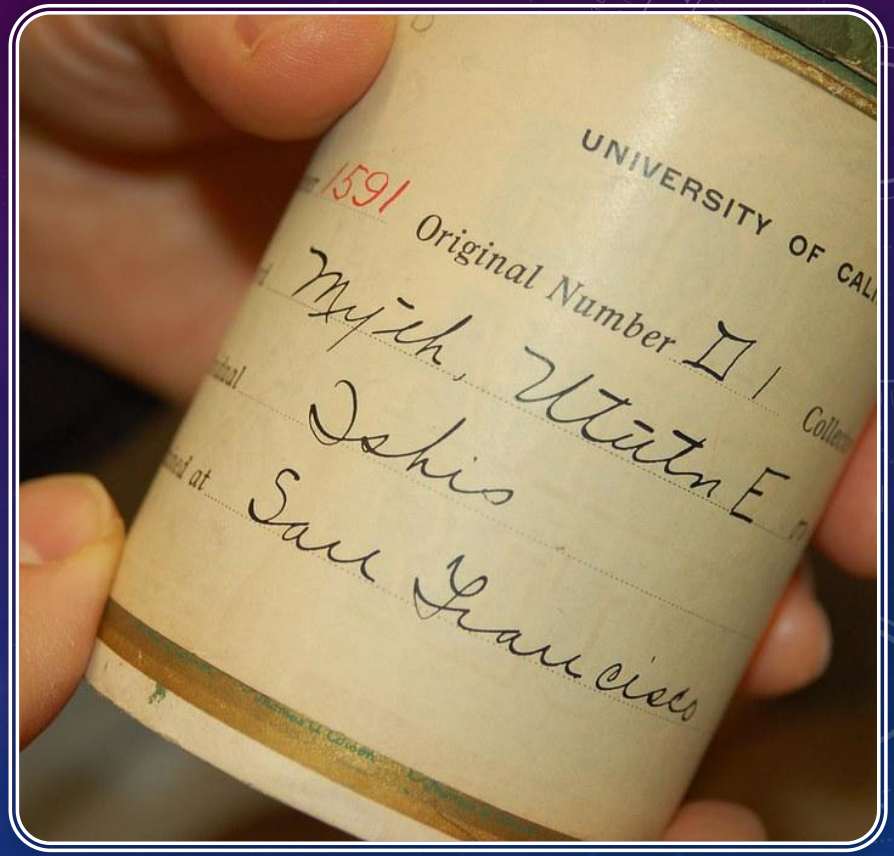
WHAT IS AN ARCHIVE?

Repository of language data

- With the aim of preserving and disseminating materials in its collections (Burke et al. 2021; Austin 2021; Kung et al. 2021)

Austin (2021):

- **Appraise** materials
- **Preserve** them longterm
- "Make their existence **discoverable**"
- Facilitate their appropriate **distribution**



Appraisal (what gets archived)

What goes into an archival collection?

- What the archives choose to collect
  - Grey literature?
  - Opportunistic vs planned collection? (e.g. scraped corpora vs results of fieldwork)
- What the depositor records
  - “Data” - about language, culture, participants, etc
  - Metadata

Archive “Scope”

Size	Archives of archives	Scope of collections	Relationship to physical archives	Institutional support
• 1 or 2 resources • ... • major regional holdings • global scope	• e.g. British Library ‘endangered archives’	• Local, regional or global focus • Only language materials (vs other cultural collections, e.g. music) • Publications, grey literature • How much historical data? • Cf. web corpora (e.g. sketchengine)	• Portal to physical archives, listing material onsite • Digital archive of physical holdings • Separate collections	• Departmental • Housed within larger body (e.g. university or museum system) • self-standing

Preservation

Are linguists creating stable, sustainable collections?

- **Sort of?? Kinda???**

- Pay attention to file formats and metadata
- BUT substantial issues with usability and what is archived
  - Software obsolescence, backwards compatibility (even in open source formats)
  - Data ~ metadata linkage

Babinski et al (2022)

How are archives preserving materials longterm?

- **For the most part, we don’t know!**
- Few have public information about CMS or infrastructure plans
  - Omeka
  - Mukurtu
  - Bespoke platforms
- Archives are chronically understaffed and underfunded!

Discoverability

Accounts and Registration

- The majority of archives are **open access**, requiring free account registration at most
  - Tiers of restriction: open-access, registration required, special permissions
  - This can be cumbersome
- Many archives appropriately have access restrictions for collections to respect the wishes of language communities and researchers
  - Not just free for all data, and that's a good thing
  - Not all archives streamline this process or provide clear contact information
- Some could only be accessed through an institutional (e.g. .edu) email
  - Some collections were password-protected

Searches

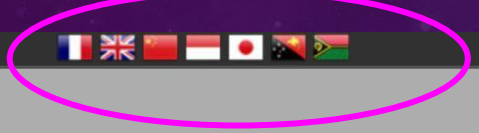
- Search engine localizations make multilingual searches difficult
- Heavy bias towards returning results in major languages
- Clearinghouses like OLAC partially remedy this problem
- But OLAC not complete, difficult to navigate

Collections are discoverable for English-medium academics…

Distribution

- How findable are collections?
- How easy are they to access?
- How downloadable are they?

ACCESSIBILITY



Interface language

PARADISEC

Home About us Our Collections Deposit Resources Support Us Donate

PARADISEC

Ostrelia hemi stap long medel blong fulap defdefren kalja mo lanwis long Pasifik. Bitim 2000 lanwis oli stap long Ostrelia, Pasifik aelan mo PNG (we igat klosap 900 lanwis i stap). Be, kasem en blong yia 2100 maet i gat plante lanwis i lus. Ol rikoding we i stap oli no save kasem evri samting we lanwis i save talemaot long hem, mo evri spesel kaen samting we wanwan lanwis i gat long hem (olsem ol singsing, ol toksave, ol storian).
PARADISEC (Pacific And Regional Archive for Digital Sources in Endangered Cultures) hem i wan organaesesen we i stap lukaotem ol rikoding mo pepa long ol defdefren lanwis. Mifala i bin mekem wan sef ples blong lukaotem ol samting ia, mo seremaot wetem ol man we i stap long rikoding, o maet sam bumbu blong ol i stap.
Mifala i bin wok tugeta wetem Vanuatu Kaljoral Senta, Iunivesiti blong Niu Kaledonia, Museum blong Solomon Aelen, Institiut of Papua New Guinea Studies mo sam man blong Rapa Nui (Ista Aelen).

hawaiisignlanguage-woodward-03...
hawaiisignlanguage-woodward-0345
To'plam
To'plam dastasi : http://hdl.handle.ne...
Sarlavha : Gavayi imo -ishora tili hujjat...
Manzil
Qit'a : Amerika

Mamlakat : AQSh
Hudud : Amerika
Manzil : Gavayi
Tillar
Til
Ism : Gavayi imo -ishora tili

Kod : ISO639-3: ot kuchi
Loyiha
ID : MDP0278
Ism : 0345-MDP0278
Tavsif : Til_Nomi: Gavayi imo-ishora til...
Loyiha haqida ma'lumot

Depozit holati : onlayn yig'ish
Moliyalashtiruvchi organ : ELDP
Aktyor
To'liq ism : Jeyms Vudvord
Vazifasi : omonatchi
Aloqa : Manoa shahridagi Gavayi univ...

Ko'proq ko'rsatish

Ichidan qidirish JILD

TANLOVINGIZNI YAXSHILANG

Shaxs

- Hilda Lopez (7)
- Jeyms Vudvord (23)
- Jeyms Vudvord (1)
- Linda Lambrecht (23)

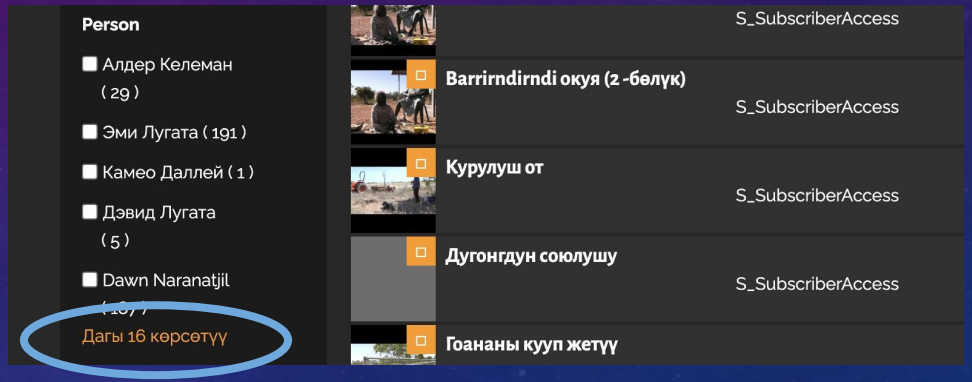
1889-1940 yillar Miki yubileyidan fotosuratlar 2004 yil
U_UserAccess

1941-1950 yillar
U_UserAccess

1941-1950 yillar Mikining yubiley fotosuratlari
U_UserAccess

1951-1960 yillar
U_UserAccess

“View 16 more” - menu buttons no longer work with Google Translate as overlay



FUNCTIONALITY

SITE CONTENT, STRUCTURE, AND DOWNLOADS

- No bulk-download (34/41 archives)
  - This makes large collections (e.g. those with 15,000+ files) inaccessible, and even smaller collections time-consuming to download
  - (Other solutions, e.g. scraping, often violate terms of use)
- File downloads can cause loss of nested file structure, used in covert metadata

Dissemination vs Curation vs Longevity

Audiences differ in dissemination needs

In-browser
Downloadable
Item by item vs linked collections

Content management systems

Help in longevity
But not always for findability
Or for bulk retrieval

(File size, connection issues)

CMS and "intermediate archiving"

GENERAL CONCLUSIONS

INTERIM SUMMARY

Austin (2021):

- **Appraise** materials :: unknown
- **Preserve** them longterm :: methods unknown, known to be heterogeneous!
- “Make their existence **discoverable**” :: variable
- Facilitate their appropriate **distribution** :: variable

DEPOSITOR LEVEL-ARCHIVE LEVEL INTERACTIONS

- Archive Level protocols create standards for individual depositors to follow
- Depositor Level practices will help uphold the consistency of inter-archive and intra-archive usability

IN ADDITION…

- Archives and depositors can make their collections more **usable**.
  - What's archived
  - Format and order of collections
  - Annotated and metadataed
  - In stable and open-source formats
- How can we consider the many different ways language data is now used:
  - Interacted with item by item (e.g. human plays recordings, listens to/watches them)
  - Re-formed or republished
  - Used by humans to make other resources in or about the language
  - Used in extracting statistical information about the language
  - Used to train computational models (e.g. for speech to text)
  - …

Some things that don't matter for humans matter a lot for computers, and vice versa

THANK YOU!