WHAT IS IN A LANGUAGE ARCHIVE AND HOW TO FIND IT

- Scale of documentation currently held and planned for the future
- Complicated role of archives (colonial or post-colonial?)
- Publicising items in collections to maximise chances of speakers finding it
- Building local collections
“it is becoming feasible to record around 500 hours of linguistic material in the course of a year or two’s fieldwork, thanks to the miniaturisation, fidelity and portability of our recording devices.” (Evans 2017: 41)

This will require

- a change in fieldwork methods
- more archives and increased capacity and efficiency of ingestion and delivery
  - 500 hours of audio and video is a terabyte or more of storage
LANGUAGE ARCHIVES

- Digital Endangered Languages and Musics Archives Network (DELAMAN)
  - AILLA (Archive of the Indigenous Languages of Latin America)
  - Alaska Native Language Archive
  - California Language Archive
  - ELAR (Endangered Languages Archive at SOAS)
  - Kaipuleohone (University of Hawai‘i Digital Ethnographic Archive)
  - Native American Languages Collection at the Sam Noble Museum of Natural History
- PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures)
- RWAAI (Lund University)
- The Language Archive (Max Planck Institute for Psycholinguistics)
- The Library of the American Philosophical Society (APS)
- World Oral Literature Project
LANGUAGE ARCHIVES

- Open Language Archives Community
  - Academia Sinica Collections
  - Alaska Native Language Archive
  - Archive of the Indigenous Languages of Latin America (AILLA)
  - C'ek'aedi Hwnax Ahtna Regional Linguistic and Ethnographic Archive
  - California Language Archive
  - CHILDES Data repository
  - COollections de COrpus Oraux Numeriques (CoCoON ex-CRDO)
  - Comparative Corpus of Spoken Portuguese
  - Endangered Languages Archive
- Kaipuleohone
- Living Archive of Aboriginal Languages
- PARADISEC
- Pacific Collection at the University of Hawai'i at Mānoa Hamilton Library
- PANGLOSS
- TALKBANK Data repository
- The Crúbadán Project
- The Language Archive at the MPI for Psycholinguistics
- The LDC Corpus Catalog
- Tibetan and Himalayan Digital Library
LANGUAGE ARCHIVES

- Organise linguistic records
- Depositing in an archive may be the first time that a linguist organises their collection
- Provide a citable form of language records
- Curate records with licenses for use
- Make records available to source communities
- Are always underresourced and so need to prioritise their activities
<table>
<thead>
<tr>
<th>Archive</th>
<th>Number of Resources</th>
<th>Distinct Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archive of the Indigenous Languages of Latin America (AILLA) (USA)</td>
<td>289</td>
<td>257</td>
</tr>
<tr>
<td>C’ek’aedi Hwnax Ahtna Regional Linguistic and Ethnographic Archive (USA)</td>
<td>1,474</td>
<td>3</td>
</tr>
<tr>
<td>California Language Archive (USA)</td>
<td>14,582</td>
<td>281</td>
</tr>
<tr>
<td>COollections de COrpus Oraux Numeriques (CoCoON ex-CRDO) (France)</td>
<td>13,437</td>
<td>217</td>
</tr>
<tr>
<td>Endangered Languages Archive (SOAS, UK)</td>
<td>89,575</td>
<td>463</td>
</tr>
<tr>
<td>Kaipuleohone (USA)</td>
<td>4,785</td>
<td>199</td>
</tr>
<tr>
<td>The Language Archive (Nijmegen, Netherlands)</td>
<td>127,985</td>
<td>433</td>
</tr>
<tr>
<td>Living Archive of Aboriginal Languages (Australia)</td>
<td>2,966</td>
<td>38</td>
</tr>
<tr>
<td>Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) (Australia)</td>
<td>17,084</td>
<td>1202</td>
</tr>
<tr>
<td>Pacific Collection at the University of Hawai’i at Mānoa Hamilton Library (USA)</td>
<td>5,292</td>
<td>749</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>277,469</strong></td>
<td></td>
</tr>
</tbody>
</table>

January 2019
OPEN LANGUAGE ARCHIVES COMMUNITY (OLAC)

http://www.language-archives.org/language/erk

Using ISO-639-3 codes for languages

OLAC resources in and about the South Efaté language

ISO 639-3: erk

The combined catalog of all OLAC participants contains the following resources that are relevant to the language:

- Primary texts
  - https://example.com/text1
  - https://example.com/text2
- Language description
  - https://example.com/description
- Other resources about the language
  - https://example.com/other

Other known names and dialect names: South Efaté, Eker, Fata, Southern Efaté

Use faceted search to explore resources for South Efaté language.

Primary texts


Language descriptions


Other resources about the language

PACIFIC AND REGIONAL ARCHIVE FOR DIGITAL SOURCES IN ENDANGERED CULTURES (PARADISEC)

» Collaborative digital research resource set up by University of Sydney, University of Melbourne & Australian National University, 2003 with the aim of preserving ethnographic (language, oral history, music) recordings made since the 1950s

funding from Australian Research Council
PARADISEC

- 1,203 languages
- 489 collections
- 22,937 items
- 273,160 files
- 9,650 hours (audio)
- 820 hours (video)
- 55 TB of files

- 30.7 TB of video files in archive
- 14.8 TB of audio files in archive
- 1.29 TB of image files in archive
- 1.16 GB of text files in archive

- January 2019
API, METADATA FEEDS

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 (1) developing consensus on best current practice for the digital archiving of language resources, and (2) developing a network of interoperating repositories and services for housing and accessing such resources.

News

OLAC joins the Linguistics Linked Open Data Cloud: The OLAC system has now been integrated with LLOD... http://search.language-archives.org/search.html?

New OLAC Page Listing Archive Submission Policies: As a service to linguists who are in search of an archive that could receive a deposit... http://search.language-archives.org/search.html?

New OLAC Search Service: In December 2010, to mark our 10th anniversary, OLAC announces a new search service... http://search.language-archives.org/search.html?

Language Archives Tutorial at LSA: In January 2009, on OLAC-endorsed LULAC on language archives was held at the winter meeting of the Linguistic Society of America... http://search.language-archives.org/search.html?

OLAC Presented at three Pacific Conferences: In 2006 and 2008, OLAC is being presented at conferences in Australia, the Philippines, and Hawaii... http://search.language-archives.org/search.html?

More news...

Documents

OLAC Standards - specify how OLAC operates

Recommendations - express consensus of OLAC members regarding language resource archiving

Tips - background information and guidance for implementers

General Information:

Overview (FAQ) - implement FAQ

Find Language Resources

Try OLAC’s new search engine at: http://search.language-archives.org/

OLAC Coverage

OLAC Archives contain over 100,000 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.

Feeds picked up by ANDS, OAI services, Open Language Archives Community (OLAC), etc.
HOW TO GET RECORDS BACK TO THEIR SOURCE?

- sending hard disks to cultural centres
- usb drives to plug into xbox or tv (perhaps not for computers)
- providing low resolution copies of archival files for delivery
- making our catalog maximally accessible and findable
- problem of providing context for a set of files
- google searching is the most likely way that speakers will find our records
ARCHIVES CRITIQUED AS INACCESSIBLE TO SPEAKERS

- “One major insight from the workshop was the discovery that at present, most language archives are not meeting the needs of most users.” Wasson et al 2018: 642

- We can all do better, however, PARADISEC has many users and they are spread through remote and urban areas.
ARCHIVES CRITIQUED AS INACCESSIBLE TO SPEAKERS

“Archives have made explicit attempts to serve this community; however, the vast majority of these attempts have been top-down in nature, inspired by the archives’ vision of the needs of these user communities.” (Dobrin & Holton 2013)

Another way of seeing this is that linguists are taking responsibility for repatriation.

Would a preferable response be the one that archives are always criticised for, locking up the data in inaccessible forms?
“The challenge is to design language archives that accommodate the needs of all user groups. A user-centered design process could help make language archives more accessible and more useful to larger numbers of users.” (Wasson et al 2016: 643)

Meantime, archives need to keep accessioning and making material available
Motu-Koitabu electoral roll - The National
https://www.thenational.com.pg/motu-koitabu-electoral-roll/
Apr 26, 2018 - I make reference to a report in this newspaper on the Motu-Koitabu electoral roll and the removal of 3000 foreign or non-native names from the ...

Motu Koitabu Assembly Act Never Recognised says Acting Chairman ...
https://emtv.com.pg/motu-koitabu-assembly-act-never-recognised-says-acting-chairman...
The Motu Koitabu Assembly Act 2007 has never been recognized by the Government since its...

Motu Koitabu Environment and Health ... - Project Detail - UNDP
https://sgp.undp.org › Projects › Project Search Results ▼
A planning grant of K1,966.67 was provided to the Motu Koitabu Group of the Village of Hanuabada to develop their project proposal. In late 1998 the group ...

OLAC resources in and about the Koitabu language
www.language-archives.org/language/kqi ▼
Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).
oai:paradisec.org.au:SAW1-018; ONLINEMotu, Koitabu: LMS Singing ...

 USERS FINDING THEIR LANGUAGE RECORDS
Searching for his language ‘Koitabu’: second page of a google search
If only you witnessed and captured the reaction in me going through the recordings at home! It is quite an amazing experience! From feeling of awe to emotion to deep excitement!

The feeling of knowing that your language has been documented or recorded in a structured way, kept safely somewhere in the world, hearing it spoken 50 - 60 years ago and by some people you haven't seen but whose names you only hear in history etc is quite incredible. It is most heartwarming to know that it is possible to sustain the life of my language. Thank you once again for the opportunity to listen to the records!

As I see it now, it is "US OR NEVER!" In about three or four Koitabu villages, the observation is that people in my peer group can understand in full but speak a little. Generation after us cannot do both. I believe that if we don't teach ourselves now and pass on, the language will be totally gone in the next 10 to 20 years.

E'ava Geita, Koitabu, PNG, 23/3/2015
ENGLISH ONLY?

- Of all the language archives listed earlier, most present English only interfaces to the collection, and English only metadata.
- The Archive of the Indigenous languages of Latin America, AILLA, provides a Spanish interface (and metadata labels).
- COollections de CORpus Oraux Numériques, CoCoON, is French only.
- Pangloss is French and English.
- PARADISEC is English only, but allows metadata in any language and has information brochures in French, Mandarin, Indonesian, Japanese, Tok Pisin and Bislama.
Archival files in PARADISEC are transcoded to compressed versions on ingest

- wav > mp3
- tif > jpg
- video > mp4

Both versions are provided in the collection
Folk Stories

Multimedia item 16 of 18

DOWNLOAD MP3

0:09 / 2:16

EAF

PLAY (0:02)

A_Transcription-txt-mos

are? ṅy cinlu cʰa bə vu cʰa bə aran cʰa ate cʰa. po? bə muŋkaha? ṅy ṅa bə don ḳy ṅame twɛʒu pantʰum.

Cinlu is a small bird, the elephant thinks, 'I am largest animal in the world.'

PLAY (0:15)

https://github.com/marcolarosa/pdsc-collection-viewer
PUBLICISING THE COLLECTION

- Publicity helps lead both to discovery of what is in the collection, and to accession of new material
- PARADISEC Facebook page has 1,933 followers
PUBLICISING THE COLLECTION

Glossopticon : https://www.youtube.com/watch?v=X2iimqq7LS8

https://glossopticon.com/vr.html
MAKING SUBCOLLECTIONS FOR DELIVERY

PARADISEC collection

- 20,911 items
- 238,977 files

Delivery to source community

- Item 1 (20 files)
- Item 2 (2 files)
- Item 3 (4 files)
- Item 4 (7 files)

Problem: no catalog accompanies the files
### FILE DUMP

<table>
<thead>
<tr>
<th>NT10-W11-A.wav.txt</th>
<th>NT6-200513-A.mp3</th>
<th>NT5-200601.wav</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT10-W11-A.txt</td>
<td>NT6-200511-A.mp3</td>
<td>NT5-200519-A.wav</td>
</tr>
<tr>
<td>NT10-W09-A.txt</td>
<td>NT6-200511-A</td>
<td>NT5-200518-A.wav</td>
</tr>
<tr>
<td>NT10-W08-BAMP.txt</td>
<td>NT5-TokelauOf-vid.mov</td>
<td>NT5-200513-A.wav</td>
</tr>
<tr>
<td>NT7-200805-1.mp3</td>
<td>NT5-Tokelau-Vid1.mov</td>
<td>NT5-200501.wav</td>
</tr>
<tr>
<td>NT7-200804-1.mp3</td>
<td>NT5-StringBand-007.dv</td>
<td>NT1-98021-A</td>
</tr>
<tr>
<td>NT7-200803-1.mp3</td>
<td>NT5-StringBand-001.mov</td>
<td>NT1-98020-B</td>
</tr>
<tr>
<td>NT7-200801-4.mp3</td>
<td>NT5-DickLauto-Vid2.mov</td>
<td></td>
</tr>
<tr>
<td>NT7-200801-1.mp3</td>
<td>NT5-200801-4.wav</td>
<td></td>
</tr>
<tr>
<td>NT6-200518.mp3</td>
<td>NT5-200801-3.wav</td>
<td></td>
</tr>
<tr>
<td>NT6-200517-A</td>
<td>NT5-200605-A.wav</td>
<td></td>
</tr>
<tr>
<td>NT6-200515-A.mp3</td>
<td>NT5-200604.wav</td>
<td></td>
</tr>
<tr>
<td>NT6-200513.mp3</td>
<td>NT5-200602.wav</td>
<td></td>
</tr>
</tbody>
</table>
PARADISEC CATALOG WRITES XML FILES

<?xml version='1.0' encoding='utf-8' ?>

<item>
  <identifier>NT1-006-006A</identifier>
  <collectionId>154</collectionId>
  <private>false</private>
  <title>Recordings in South Efate</title>
  <description>Stories of Longtatalof, and Kata</description>
  <originDate>1997-11-12</originDate>
  <archiveInfo>
    <archiveInfo>
      <dateOfRecording>1997-11-12</dateOfRecording>
      <dateDigitised>2001-01-01 00:00:00 +1100</dateDigitised>
      <ingestNotes/>
      <metadataImportedOn/>
      <metadataExportedOn/>
      <tracking>1 CD given to the Yanuatu Cultural Centre</tracking>
    </archiveInfo>
    <archivalInfo>
      <archivalInfo>
        <recordCreated>2012-03-27 20:08:00 +1000</recordCreated>
        <recordLastModified>2016-03-29 02:13:43 +1100</recordLastModified>
        <adminAccess/>
      </archivalInfo>
      <citation>Nick Thieberger (collector), Nick Thieberger (depositor), Kalsarap Namaf (speaker), Iokapeth (speaker)</citation>
    </archivalInfo>
  </archiveInfo>
  <agents>
    <agent role="depositor">
      Nick Thieberger
    </agent>
    <agent role="recorder">
      Nick Thieberger
    </agent>
    <agent role="speaker">
      Kalsarap Namaf
    </agent>
    <agent role="speaker">
      Iokapeth
    </agent>
  </agents>
  <collection>University of Melbourne</collection>
  <operators>Corin Bore</operators>
  <dataCategories>
    <category>primary text</category>
    <category>sounds</category>
  </dataCategories>
  <discourseType>
    <agent>
      <language>Bislama - bis</language>
      <language>Telep - 1pm</language>
    </agent>
    <country>Vanuatu - VU</country>
  </discourseType>
  <files>
    <file name="NT1-006-006A.wav">
      <mimeType>audio/x-wav</mimeType>
      <fileSize>331 MB</fileSize>
      <duration>00:32:47.569</duration>
      <bitrate>1.35 MB</bitrate>
      <sampleRate>44100 Hz</sampleRate>
      <channels>2</channels>
    </file>
    <file name="NT1-006-006A.mp3">
      <mimeType>audio/x-wav</mimeType>
      <fileSize>331 MB</fileSize>
      <duration>00:32:47.569</duration>
      <bitrate>1.35 MB</bitrate>
      <sampleRate>44100 Hz</sampleRate>
      <channels>2</channels>
    </file>
    <file name="NT1-006-CAT-PDSC_ADMIN.xml">
      <mimeType>text/xml</mimeType>
      <fileSize>331 MB</fileSize>
      <duration>00:32:47.569</duration>
      <bitrate>1.35 MB</bitrate>
      <sampleRate>44100 Hz</sampleRate>
      <channels>2</channels>
    </file>
  </files>
</item>
DOWNLOADER

- Harvests individual catalog xml files and creates a static html page catalog of just that assemblage of items
- We had previoulsy built a javascript viewer for items in PARADISEC and re-used that to be the browser
- After testing, we added new functionality to group items by genre

- http://www.language-archives.services/about/data-loader
ADDING GENRE AS A METADATA TERM

- Adapting the catalog to suit a particular requirement
- Comments field used to group items by genre
- Captured by the dataloader for delivery on raspberry pi
RASPBERRY PI

- Delivery of archival files avoiding bandwidth cost
- Raspberry Pis are low powered webservers
  - wifi transmitter
  - USB drive stores the data

http://www.language-archives.services/
https://github.com/marcolarosa/pdsc-librarybox-installer
Mobile phones – 80% coverage in most Melanesian countries

Port Vila, Vanuatu, ‘Internet on your phone’, 2012
ACCESS VIA MOBILE PHONES

- Tested in Vanuatu, Tahiti, Western Desert (Australia)

Elisa Alphonse in Erakor village

Kalonuk Albert & Gino Joseph in Erakor village
ARCHIVES PROBABLY REQUIRE LOCAL INTERMEDIARIES

- Archives should have the preservation of primary records as their main goal.

- If someone can navigate google then they are probably going to find records in their language and can then negotiate the interface of some, if not all, current language archives.

- For those who are unable to search the web, local agencies can present archival information in more relevant ways.
  - language centres, cultural centres

- These are where the raspberry pi can sit and broadcast for phone users.
REFERENCES


Australian Research Council – LIEF grants LE110100142, LE0560711, LE0453247

ARC DP0450342, DP0984419, FT140100214, CE140100041