Going forward with language archives
Given the relentless
- entropy that degrades our field recordings, and
- innovation that makes the technology we have used to capture them obsolete within a decade

We know that
- those recordings are just as endangered as the languages they document, unless
  - they are entrusted to archives for long-term preservation

So why then is the following the case?
- The vast majority of field recordings remain unarchived
What is holding linguists back?

- In order to realize the long-term benefit, there are a number of short-term costs:
  - “I will have to learn how to do archiving.”
  - “I will have to do a lot of work to organize my recordings and add the metadata.”
  - “I need to do more transcription and annotation before my materials are ready.”
  - “If I let the material go, somebody may publish on them before I do.”
- And so archiving gets put off until a better time in the future—which may never come
The AARDVARC hypothesis

- The initial hypothesis in the AARDVARC proposal:
  - We could incentivize more archiving by using automation to break the transcription bottleneck

- A more refined hypothesis has come out of the series of AARDVARC workshops:
  - We could increase archiving by leveraging automation wherever possible, both
    - To add incentives for archiving, and
    - To remove disincentives
Going forward, the future of language archives is "automated services"

### Leveraging automation

<table>
<thead>
<tr>
<th>By offering ...</th>
<th>An archive can ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automated ingest services</td>
<td>Remove obstacles to submission</td>
</tr>
<tr>
<td>Automated presentation services</td>
<td>Provide incentives for early submission</td>
</tr>
<tr>
<td>Automated annotation services</td>
<td></td>
</tr>
</tbody>
</table>
An example for automating ingest:
Language documentation at SIL International

- We have good software tools for Lang Doc and a well-used digital archive with on-line submission
  - But primary recordings are not being archived
- SIL’s archive already has these incentives in place:
  - The peace of mind of long-term preservation
  - A citable “publication” that others can access
  - Management of graded access to sensitive content
- But these are eclipsed by a huge disincentive:
  - There is too much learning and work involved in turning a compiled collection into an archived corpus
The three basic tasks of Lang Doc

“Language Documentation is concerned with compiling, commenting on, and archiving language documents.”
— Himmelmann 1998

1. **Compile** a sample of recordings of a full range of speech event types

2. **Comment** on those recordings
   - E.g., transcription, translation, discussion, situational context, informed consent to share

3. **Archive** the complete corpus of recordings and commentary with an institution that will provide long-term preservation and access
The status quo for SIL tooling

- We have a great tool for compiling and commenting
  - **SayMore**: “Language Documentation Productivity”
    - Organizes all the files and their associations
    - Records metadata on sessions and people
    - Tracks progress on commenting workflow
    - Supports respeaking, transcription, translation
    - Download v. 3.0 at [http://saymore.palaso.org/](http://saymore.palaso.org/)

- But it falls short of supporting the entire enterprise
  - Users are on their own to figure out how to archive their whole collection
Automating ingest involves both preparation of the submission package and intake into the archive
- Enhance SayMore to create archive submission package
- Use API on the digital archive to automate submission

The value proposition to the linguist should be:
- “You can archive your corpus at the push of a button!”

Requirements:
- A single command causes a SayMore project to be packaged as a corpus and submitted to the archive
- The archive submission package is known to be complete and well-formatted
Reasons for returning a submission

- The metadata for the project, the sessions, or the participants is incomplete
- There is no introductory document describing the project and its methods
- There are no “Table of contents” documents listing all the sessions and all the participants
- There are materials marked for release to the public that lack informed consent to share
- There are participants who have not given consent for public identification and have not been anonymized
- There are files not attributed to any participants or in formats that are not accepted by the archive
Specifications for the updates to SayMore

- Archivists have identified information that is absent
  - Some metadata fields that are missing in SayMore
  - No slot in the project for an Introduction document
  - No “Requests anonymity” check box for participants

- And a “Preflight for archiving” function is needed which:
  - Warns of a missing Introduction
  - Identifies every missing obligatory metadata element
  - Identifies every file that is not attributed to any participant
  - Identifies every file in a format not accepted by the archive
  - Identifies every session marked for public release that is missing informed consent to share
Specifications for “Archive now” button

- Update the automatically generated “tables of contents”
- Generate and insert the “preflight” report for the curator
- Organize the sessions into collections by access level, while anonymizing as needed
- Place the key to anonymization in a curators-only folder
- Generate the corpus metadata record as a METS package
- Bundle the corpus contents into bitstreams that are ZIP files of up to 1 Gigabyte each
- Use SWORD API on the DSpace repository to automate submission of the METS package and all the bitstreams
Another example:
Language Preservation 2.0 and Aikuma

- An NSF grant project by Steven Bird ([http://lp20.org](http://lp20.org))
  - Language Preservation 2.0: Crowdsourcing Oral Language Documentation using Mobile Devices

- The centerpiece is Aikuma
  - An Android app
  - Community members make recordings
  - Share and vote via Wi-Fi router w/ storage
  - Two-button app for time-aligned respeaking and oral translation
  - Automated upload to the Internet Archive
Automating presentation services

- **Status quo**
  - A linguist deposits a corpus to an archive
  - The corpus becomes discoverable through OLAC
  - A user downloads materials to explore on own system

- **Envisioned future**
  - Upon ingest, the archive automatically creates a web space that presents the corpus content to users
  - An immediate benefit of automated deposit is simultaneous presentation of materials to language community members, scholars, and the public
An example: EOPAS as a good starting point

- Ethnographic E-Research Online Presentation System, from School of Language and Linguistics, University of Melbourne

Making thatch; Efate, South (erk); Vanuatu

- "I clean them, they call them pins."
- "I clean the pins until it is done."
- "And then I’ll sew the thatch."

Depositor: Nick Thieberger

Recorded: 2011-02-15 00:00:00 UTC
Removing the transcription bottleneck

- An open source project (http://www.eopas.org)

- Current functionality
  - Starts with transcription to anchor the display
  - Adds interlinear analysis and translation as available

- Additionally needed functionality
  - Handle recordings with no transcription
  - Incorporate aligned respeaking when available
  - Incorporate oral translation when written not available
  - “Keyword spotting” for phonetic search over recordings
Status quo

- Linguists perceive completion of transcription (and other annotation) as a prerequisite for archiving
- Linguists typically attack this problem by themselves
- They do not use state-of-the-art automated annotation tools since they aren’t easily installed
  - speech activity detection
  - speaker diarization (i.e., segmenting into turns with speaker id)
  - automatic transcription of oral translations in major languages
  - machine learning of models for language-specific annotation
Automating annotation services

- Envisioned future
  - Archives provide for processing of deposited materials with state-of-the-art automated annotation tools
  - An immediate benefit of archival deposit is access to these automated annotation tools
  - A further benefit is that other web users (e.g., language community members, citizen scientists) can use the tools to help with transcription and annotation
  - Archive deposits are progressively enriched via stand-off annotations attributed to the annotator so that absence of annotation need no longer delay archiving
The Language Application Grid: A Framework for Rapid Adaptation and Reuse

Vassar, Brandeis, CMU, Linguistic Data Consortium

The Grid consists of:

- Data services—Provide access to corpora
- Processing services—Provide access to natural language processing (NLP) tools
- Composition of services—Creating workflows to run data through one or more processes
- An archive could provide services by joining the Grid
So what’s in the future of digital language archives?
- Automation!

Archives will make the transition from being just the final stop for long-term preservation to becoming an early stop for essential services now and in the future:
- Automated services to break the ingest bottleneck
- Automated services to break the annotation bottleneck
- Automated services to present archived language documentation to its potential users in such a way that it meets their needs