Language Documentation & Archiving

Heidi Johnson
The Archive of the Indigenous Languages of Latin America (AILLA)
The University of Texas at Austin
Acknowledgements

Language Digitization Project Conference 2003

EMELD Working Group on Resource Archiving:

Gary Holton, ANLC
Heidi Johnson, AILLA
Nick Thieberger, PARADISEC
Gary Simons, SIL International
Wallace Hooper, Indiana University
Susan Hooyenga, University of Michigan

http://emeld.org
A little history

- Boasian tradition: grammar, dictionary, collection of texts
- Linguists gave field materials to museums & libraries, e.g. Smithsonian. Seeking a permanent home for endangered language materials.
- M & L not really able to preserve recordings, other than by storing them in a cool dark place.
History, cont.

Anything that can be published was & is a 
**distillation** - the product of analysis. 
Secondary/tertiary resources.

Hitherto no feasible means of preserving 
OR publishing primary materials.

The new millenium: digital archives can 
preserve and/or publish anything.
What is an archive?

**Archive:** a trusted repository created and maintained by an institution with a *demonstrated commitment to permanence* and the long-term preservation of archived resources.

**Collection:** the body of documentary materials created by researchers and native speakers. Serves as the basis for research & education. Will be deposited in an archive.
Why should you archive?

- to preserve recordings of endangered/minority languages for future generations.
- to facilitate the re-use of primary materials (recordings, databases, field notes) for:
  - language maintenance & revitalization programs;
  - typological, historical, comparative studies;
  - any kind of linguistic, anthropological, psychological, etc. study that you yourself won't do.
More reasons to archive

- to foster development of both oral and written literatures for endangered languages.
- to make known what documentation there is for which languages.
- to build your CV and get credit for all your hard work.
Archiving is a form of publishing

- Even if the resources are restricted, the metadata is public.
- Get credit for fieldwork in the early stages: list Archived Resources on your CV.
- Cite data from archived resources.
- Give consultants proper credit for their work and their creations.
Citing archived resources

What should you archive?

- Recordings of discourse - audio and/or video - in as wide a range of genres as your community employs.
- Always get permission for everything:
  - recording
  - archiving
  - excerpting, publishing, etc.
Things you should archive

- public events: ceremonies, oratory, dances, chants
- narratives: historical, traditional, myths, personal, children's stories, ...
- instructions: how to build a house, how to weave a mat, how to catch a fish, ...
- literature: oral or written, poetry, any creative work
- conversations: anything that's not gossip or too personal, e.g. what we did last spring festival
More things you should archive

- transcriptions, translations, & annotations of recordings
- field notes, elicitation lists, orthographies - anything other people might find useful
- datasets, databases, spreadsheets - your secondary (unpublishable) materials
- sketches of all kinds: grammar, ethnography
- photographs
Things you should not archive

◆ Anything that would cause injury, arrest, or embarrassment to the speakers.
  
  Example: Pamela Munro's interviews with Zapotechs in L.A. about entering the U.S. illegally.

◆ Sacred works with highly restricted uses. But talk to people about safe ways to preserve such works, if they want.
How should you manage your collection?

Corpus management rule #1: Label everything you produce with RUTHLESS CONSISTENCY.

Corpus management rule #2: Set up a system before you leave & test it along with your equipment. (Tape your friends and relatives to try things out.)
1. Find an archive & get their guidelines

- DOBES, for their grant recipients: http://www.mpi.nl/DOBES
- Regional archives: AILLA, ANLC, PARADISEC, others? (See AILLA's Links page)

Note: it's not either/or, it's both/all.
If there isn't one, write to any one of us, we'll help you.
2. Identify your archival objects

- Not necessarily the same as a file or a tape.
- Language documentation materials typically come in related sets, or bundles.
- Be aware of relations among materials as you create them so you can label them correctly and keep them together.
Relations among items

- derivation: e.g. a transcription is derived from a recording
- series: e.g. a long recording that spans several tapes/discs
- part-whole: e.g. video & audio recordings made simultaneously of the same event
- association: (fuzzy) e.g. photographs of the narrator of a recording, commentaries
3. Labelling field materials

Nothing could possibly be more important than labelling every single item you produce - track, tape, disc, notebook, file slip, digital file, photograph - with RUTHLESS CONSISTENCY.
Example 1: AILLA resource ID

ZOH001R040I001.mp3

- ZOH = language code
- 001 = deposit number (first deposit)
- R040 = 40th resource in that deposit
- I001 = 1st item in that resource
- .mp3 = what kind of file

If you have an archive, write and ask them for labelling guidelines.
Example 2: participant initials plus a media type code

- gsm1_au1  audio part 1
- gsm1_au2  audio part 2
- gsm1_db   shoebox interlin of the audio
- gsm1_tx1  text, misc notes
- gsm1_ph1  photo of Germán
Example 3: label by media unit, recordings are primary

- md1t1 - minidisc 1, track 1
- md1t1.db - shoebox database for that text
- nb1 - field notebook 1
- ds19.xls - spreadsheet dataset (e.g. verb roots)
Metadata I

Catalog information for digital resources.

Supports

- archive & collection management
- protection of sensitive materials
- searching
- use of resources by many people
- proper citation of archived resources
Metadata II : Minimum info

- Speakers' full names (plus alias if you want to anonymize in text).
- Language: Be specific! Zoque of San Miguel Chimalapa, Oaxaca, Mexico.
- Date of creation: YYYY-MM-DD. Use the primary (recording) date for the bundle.
- Place of creation: Be specific: village, state, country, or river valley, region, country…
- Access restrictions & instructions, if necessary.
- Genre keyword: dependent on choice of schema.
Choose either IMDI or OLAC schema. If you have an archive, use the one they tell you.

**LABEL** every metadata entry with the same label you use for the resource. List every related item in the metadata.
IMDI: www.mpi.nl/IMDI

Session bundle = resource
- Title, date, place, description
- Depositor (you): contact info
- Project: name, director, sponsor, etc.
- Participants: role, demographic data, contact
- Resources: provenance, formats, relations, etc.
- Content: context, genre, narrative description, etc.
- References: relevant publications
OLAC:
www.language-archives.org/

Archival object definition is up to you
- Contributors / creators
- Title, date, description
- Resource info: formats
- Relation to other objects
- Subject - linguistic subfield
- Type.linguistic = genre
Corpus management tools

- From MPI: IMDI Browser & IMDI Data entry.
- I have a Shoebox 2.0 template that needs porting to Shoe 5.0 (?)..
- Someday, we'll do a Filemaker Pro one.
- Otherwise, use any database or spreadsheet or Word template and create your own.
Intellectual property rights

Define a policy concerning IPR and develop a consistent practice for obtaining consent, e.g., forms and/or recorded statements.

Learn how to talk to your consultants about IPR.

Ask other researchers who have worked in your region or language community.

Note the IPR status of each resource and each item in the metadata.
## Formats

<table>
<thead>
<tr>
<th></th>
<th>Text</th>
<th>Audio</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>archival</td>
<td>tiff / XML</td>
<td>wav 44.1/16</td>
<td>mp2</td>
</tr>
<tr>
<td>presentation</td>
<td>pdf / html</td>
<td>mp3</td>
<td>??</td>
</tr>
<tr>
<td>working</td>
<td>ms / MS Word</td>
<td>minidisc</td>
<td>??</td>
</tr>
</tbody>
</table>
Archive-quality formats are:

- non-proprietary; that is, the encoding is in the public domain;
- supports forward migration to new formats;
- portable, re-useable, repurposeable;
- best possible reproduction of the original.
When should you archive?

As soon as you get back from the field:
- to prevent accidental damage or loss;
- to get back handy presentation formats;
- to build your CV even before you are ready to publish results.

If not then, as soon as possible.

At the very least, mention your data and an archive in your will.
Archive your data

- We encourage you to archive recordings ASAP and add transcriptions, translations, annotations, etc. later.
- Secondary materials are generally reproducible; the primary recordings are not!
- Students should password-protect their data until they finish their theses.
Useful websites

- DELAMAN: http://www.delaman.org/
- IMDI: http://www.mpi.nl/ISLE
- OLAC: http://www.language_archives.org
- EMELD: http://emeld.org
- AILLA: http://www.ailla.utexas.org/links.html
- Write to me: ailla@ailla.org