Corpus Management 101:
Creating archive-ready language documentation

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

Who should archive?

- Speakers, linguists, anthropologists, …
- Anyone who wants the language documentation materials that they produce to survive and remain useful for generations to come.
- In other words: YOU.

Where should you archive?

Definitions and distinctions:

- **Archive:** a trusted repository created and maintained by an ***institution*** with a ***demonstrated commitment to permanence*** and the long-term preservation of archived resources.
- **Language documentation corpus:** the collection of documentary materials created by researchers and native speakers.

What should you archive - I

- Recordings, both audio & video:
  - public events: ceremonies, oratory, dances…
  - 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 - any creative work
  - conversations: anything that's not too personal

What you should archive - II

- Secondary (derived) materials:
  - transcriptions, translations, & annotations of recordings
  - field notes, elicitation lists, orthographies
  - datasets, databases, spreadsheets
  - sketches, e.g. grammar, ethnography
- Photographs
- Otherwise unpublished or out-of-print articles

What you should archive - III

- Teaching and learning materials:
  - primers – children's readers
  - calendars, posters, etc.
  - illustrated dictionaries, encyclopedia
  - curriculum designs
  - anything that other people might find inspiring and useful in their own programs.

What you should NOT archive

- Anything that could cause injury, arrest, or embarassment to the speakers, e.g.:
  - Pamela Munro's interviews with Zapotecs in L.A. about entering the U.S. illegally.
  - Gossip that hasn't aged enough (ancient gossip becomes history & narrative)
- Sacred works with highly restricted uses.

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.
- Restrict access to works in progress.
- Add transcriptions, annotations, etc. later.

Why should you archive? I

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

Why should you archive? II

- 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 speakers proper credit for their work and their creations.

Citing archived resources

Sánchez Morales, Germán. (1994). "Satornino y los soldados." [audio] Heidi Johnson, (Researcher.) [online] ZOH001R010. http://www.ailla.utexas.org: Archive of the Indigenous Languages of Latin America. Access=public.

How to build an archive-ready corpus I

- Rule #1: Label everything you produce with **RUTHLESS CONSISTENCY.** If I don't know what it is, I can't archive it.
- Rule #2: Get in touch with your friendly local archive and ask them to help you.
- Rule #3: Test your system before you leave: equipment, catalog method, labels.

How to build an archive-ready corpus II

- Define a policy concerning IPR and develop a consistent practice for obtaining consent, e.g., forms and/or recorded statements.
- Always get permission for everything:
  - recording
  - archiving
  - excerpting, publishing, etc.
- Learn how to talk to your consultants about IPR.

Labelling I : recordings

- Audio - record a "header" with basic information, in a contact language – English, Spanish...
  - Your name, speakers' names
  - Date & place
  - Name of the language
  - Brief statement of genre and/or title of work.
- Video - go Hollywood: use a clapboard with basic info written on it.

Labelling II: media and files

- Decide on the fundamental organizing theme for your labelling system:
  - media, e.g. CDs, notebooks
  - consultants' names or initials
  - languages/dialects
  - linguists' names or initials
  - genres, e.g. wordlists, narratives, …

Labelling III: related items

Language documentation materials typically come in related sets, or *bundles*:

- recording of a narrative + interlinear text + revised translation + commentary
- interview + photographs
- recorded elicitation session + field notes

Labelling IV: types of relations

- derivation: a transcription is derived from a recording
- series: a long recording that spans several media (cds only hold 700 mb)
- part-whole: video & audio recordings made simultaneously of the same event
- association: (fuzzy) photographs of the narrator of a recording, commentaries

Labelling V: Example: 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
- Supports our administrative needs: many languages, process one deposit at a time.

Labelling VI: media object is primary

Facilitates keeping track of things in the field. File extensions identify type of item.

- cd1t1.wav - cd 1, track 1
- cd1t1.db - the shoebox interlinear database
- cd1t1.doc - a word doc w/notes about cd1t1
- ds19.xls - spreadsheet dataset (verb roots)
- ds5.db - shoebox dataset (deictics)
- nb1 - field notebook (paper object)

Corpus catalog/Metadata I

- Catalog information for digital resources is called *metadata*.
- Metadata supports:
  - keeping related items together
  - protection of sensitive materials
  - searching for the thing you want
  - use of resources by many people
  - proper citation of archived resources

Metadata II : Minimum info

- Creators' full names: you and the speakers.
- Language: be specific.
- Date of creation: YYYY-MM-DD.
- Place of creation: be specific.
- Access restrictions, and any special instructions concerning future uses.
- Genre keyword, e.g. narrative.

Metadata III : Additional info

- Project info: name, director, sponsor, etc.
- Participants' roles (e.g. narrator), demographic data, contact info
- Resource info: provenance, formats, etc.
- Content info: descriptions of context in which created, content – the more detail here, the better for the long term.
- References: relevant publications

Metadata IV

Two recommended (interoperable) schemas. Choose either as your base and extend to suit your needs.

- OLAC – Open Language Archives Community – http://www.language-archives.org
- IMDI – International Standards for Language Engineering Metadata Initiative – http://www.mpi.nl/IMDI

Corpus management tools

- IMDI Browser & IMDI Data entry.
- AILLA's Shoebox 2.0 & 5.0 templates.
- Any database or spreadsheet or Word template that you create.
- A looseleaf binder with a standard (xeroxable) form.

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/
- Write to me: ailla@ailla.utexas.org