Archives for Language Documentation Resources

Heidi Johnson

The Archive of the Indigenous Languages of Latin America

The University of Texas at Austin

INALI, México D. F., 2 junio 2005

Preamble

Archives are the best hope of preserving and maintaining access to irreplaceable resources in endangered languages.

A little history

- Boasian tradition: grammar, dictionary, collection of texts
- Linguists gave field materials to museums & libraries, e.g. Museo de Antropología. Seeking a permanent home for endangered language materials.
- But, 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?

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

Why we need archives

- To preserve recordings of endangered languages for future generations
- To facilitate re-use of primary materials for:
  - language education programs;
  - typological, historical, comparative studies;
  - anthropological, psychological, botanical, etc. research
- To foster development of oral and written literatures in endangered languages
- To facilitate collaborative research and education.

Advantages of digital archives

Digital media in standard formats:

- Can be migrated forward over time to keep them useful forever;
- Can be copied, served over the web, and transferred to analog media such as cassette tapes;
- Can be re-used: a word list can be turned into a multimedia learning dictionary;
- Can be transferred from one institution to another on portable external hard drives.

What is in a digital archive

- Recordings: audio & video
- Texts: scanned manuscripts (notebooks, file slips, old articles and books)
- Texts: digital (e.g. Word), converted to standard formats (e.g. ascii text)
- Databases: Excel, Shoebox, etc. converted to standard formats
- Photographs: scanned or digital

Disadvantages of digital archives

- Equipment is expensive and must be up-graded every 3-5 years;
- Computational environment must be maintained by skilled professionals;
- If a digital file is mis-labelled and/or has no metadata, it may be impossible to figure out what it is.

Establishing an archive: the institutional context

Find the most stable institution that you can to host the archive.

Host institution can provide:

- Technical infrastructure and support
- Permanent full-time staff
- Credibility: an established reputation
- Connections to others working on similar projects to share resources.

Establishing an archive: the technological context

Really beyond the scope of subject-area specialists like linguists:

- Servers that can manage terabytes of multimedia resources.
- Network capacity to serve a national & international user community.
- Constantly changing technology requires experts to keep up.

Establishing an archive, cont.

Technology problem is a major reason to ally your archive with a bigger project, such as:

- La red academica?
- Internet 2:
  - http://www.noc-internet2.unam.mx/
- Biblioteca Digital Universitaria de la DGSCA:
  - http://www.bibliodgsca.unam.mx/

Mission statement

Define clearly

- the scope and scale of the collection (e.g. Mexican/Mesoamerican/Latin American languages);
- where the resources will come from (e.g. legacy materials from researchers, new materials from community projects);
- who will be the archive's primary users.

Archive management: Personnel

- General manager
- Technical manager
- Metadata specialist
- Translator
- Digitizers

Archive general manager

Should be a subject area specialist: a linguist with ties to the documentary linguistics field, and preferably with fieldwork experience.

- overall direction, oversight of other staff
- fund-raising
- acquisition of new materials
- training workshops in corpus management, archive establishment
- participate in DELAMAN (Digital Endangered Languages and Musics Archive Network)

Archive technical manager

Most likely someone from Information Sciences.

- in charge of technical standards: audio & video & text digitization;
- maintains archive software: database and website
- stays current with changes in digital technologies
- training workshops in recording & digitization
- supervises digitizers, on and off-site

Archive metadata specialist

Someone from Information Science or a linguist or anthropologist.

- Choose and customize a metadata schema.
- Participate in international metadata consortia (OLAC, IMDI, Dublin Core)
- Help depositors to develop metadata for their resources.
- Enter and review metadata for archived materials.

Archive translator

- There is a wealth of information about digital archiving, digital media, storage technologies, etc. in English that needs to be translated into Spanish.
- Metadata and other archive documents may also need to be translated into English or other languages.
- Supervise translators working in indigenous languages.

Digitization personnel

- Any patient, detail-oriented person can be trained to do this work.
- Warning! Digitizing is very boring, but it requires careful attention to detail at every step. It is very slow work.
- Let other institutions and depositors help to digitize their materials according to archive specifications. Share the work!

Define management policies

Define procedures and policies for:

- acquisition of materials, including a triage strategy for prioritizing the digitization schedule;
- dissemination of materials, including access restrictions, interface languages, etc.;
- quality assurance;
- tracking digitization standards and forward migration to new digital formats;
- disaster recovery - backups, mirror sites, etc.

Archive Management - Readings

- OAIS Reference Model for Digital Libraries
  - http://www.ccsds.org/documents/650x0b1.pdf
- EU-US Working Group on Spoken-Word Audio Collections
  - http://www.dcs.shef.ac.uk/spandh/projects/swag/
- OLAC documents:
  - http://www.language_archive.org
- DELAMAN archives:
  - http://www.delaman.org

Archive management: Overview of components

- Define the archival object: what you will preserve and how you will sort things.
- Design identifiers to label these objects.
- Choose a catalog system - called a metadata schema.
- Decide which formats to preserve and present to users.
- Determine policies for intellectual property rights.
- Develop access methods, e.g. a website.

The archival object I

Language documentation resources tend to come in **sets** of related items. At AILLA we call the whole set an **archive resource**.

- A recording with text files (transcription, translation, annotations) and photographs;
- A thesis linked to lots of audio examples;
- A Kuna poem with a version in Spanish.
- Format variations for every resource (more later).

The archival object II

- Relations among objects must be preserved by the archival method and documented in the metadata.
- It's helpful if the identifiers encode the relations, to make it easier for users to put a resource set back together.

Relations among objects

- 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

AILLA's Content Type

(Non-standard metadata element.)

primary_text
translation
interlinearization
annotation
sample
guide
transcription
transcription&translation
illustration
commentary
photograph
table of contents

Object identifiers I

Define what constitutes an archival object and be consistent in applying your definition.

1. digital objects correspond to original media: Suárez tape 1 + notebook 1, part 1 (especially good if you are also archiving the originals)
2. digital objects correspond to documentary events and/or some notion of intellectual content: "El nahual", recording + text

Object identifiers

Identifiers should support:

- retrieval of the original (analog) medium if these are preserved in the same archive;
- matching an object to related objects;
- correct citation of archived resources;
- correct relocation of the object in the archive if it gets misplaced.

Object identifiers at AILLA: I

Primary sort is by **language** (using the 3-letter Ethnologue code):

- ZOH001R040I001.mp3
  - ZOH = language code
  - 001 = deposit number (first deposit)
  - R040 = 40th Resource in that deposit
  - I001 = 1st item in that resource
  - .mp3 = file format extension

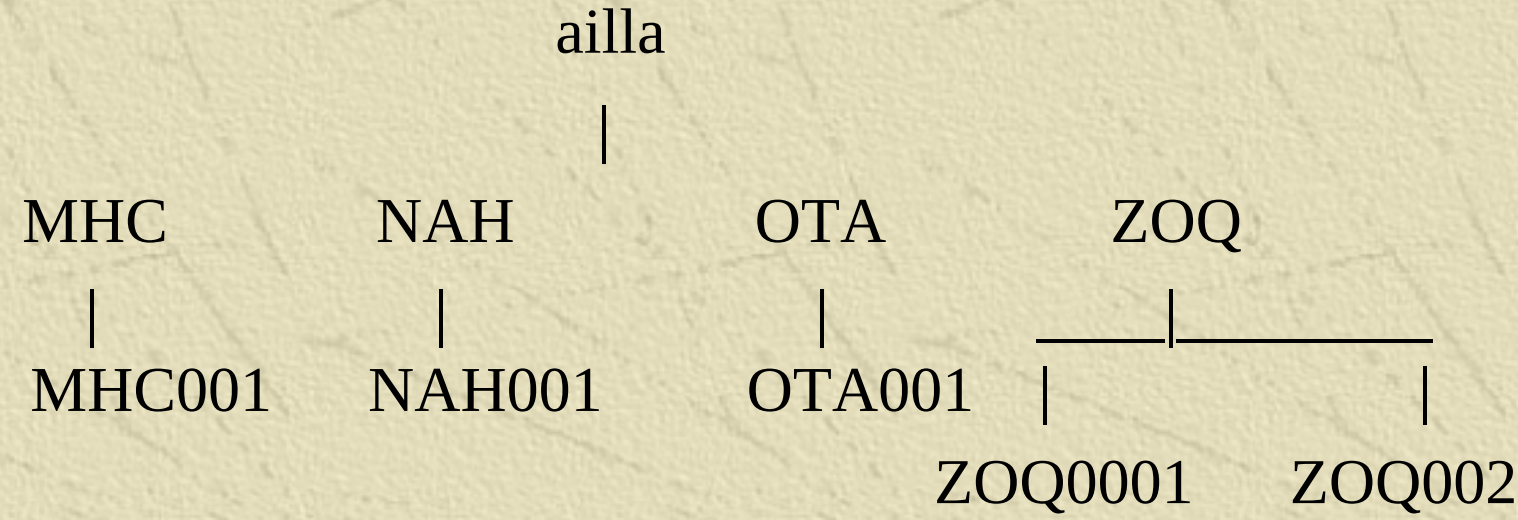
Object identifiers at AILLA: II

Files comprising resource ZOH001R020:

- ZOH001R020I001.mp3 zoque audio
- ZOH001R020I001.wav zoque audio
- ZOH001R020I001-sa.mp3 zoque audio 1-min sample
- ZOH001R020I002.mp3 spanish audio
- ZOH001R020I002.wav spanish audio
- ZOH001R020I001.txt interlinear text orig. format
- ZOH001R020I001.pdf text download format

Identifiers and archive organization

AILLA identifiers facilitate the organization of the archive file system by language and by deposit number:



Identifiers and archive organization

Other archives use different organizational criteria:

- DoBeS: organized hierarchically by project and type of data
- PARADISEC: organized by depositor/creator (identifiers start with initials)
- Existing (non-digital) archives may use existing catalog system for new digital versions.

Object identifiers, last word

- Changing identifiers is very tedious and risky work!
- Think long and hard about your system, to be sure that it reflects:
  - the kinds of materials that you will archive
  - your workflow
  - your institutional/technological context
- ***Test your system on a pilot set of resources.***

Metadata: catalog information for digital resources

- Information for administration, content description, resource description, and rights.
- Metadata schema should be OLAC-compliant.
- Best practice is to adopt and customize an existing schema (OLAC, IMDI) to maximize inter-operability.
- Be an active participant in the international language archive community.

Metadata II

Catalog information supports:

- archive management
- protection of sensitive materials
- searching
- use of resources by many people
- proper citation of archived resources

Metadata III

Two compatible schemas for language resources:

1. OLAC: The Open Language Archives Community. (Univ. Pennsylvania) http://www.language-archives.org
2. IMDI: International Standards for Language Engineering Metadata Initiative (MPI-Nijemegen) http://www.mpi.nl/IMDI

Metadata IV : minimum info

- Speakers' full names.
- Language: Be specific: Zoque of San Miguel Chimalapa, Oaxaca, Mexico.
- Date of creation: YYYY-MM-DD. Use the primary (recording) date for the set.
- Place of creation: Be specific: village, state, country, or river valley, region, country…
- Access restrictions & instructions, if necessary.
- Genre keywords, e.g. narrative, word list

IMDI

Session bundle = resource

- Title, date, place, description
- Depositor: contact info
- Project: name, director, sponsor, etc.
- Participants: name, alias, role, demographic data
- Resources: provenance, formats, relations, etc.
- Content: context, genre, narrative description, etc.
- References: relevant publications

OLAC

Archival object definition is up to you

- Contributors / creators
- Title, date, description
- Resource info: formats
- Relation to other objects
- Subject - linguistic subfield, e.g. phonology
- Type.linguistic = genre, e.g. discourse, lexicon

Example of AILLA's metadata (based on IMDI)

ID: ZOH001R020

Title Satornino 'i Soldao'øyti'

Date 1994-07-03

Place Fortín de las Flores, Veracruz, Mexico

Description: Satornino and the Soldiers. This is a local hero story, first told by the hero himself, Satornino. My consultant, German Sanchez, learned it from his father, Agripino Sanchez, who learned it from Satornino. The story takes place in San Miguel Chimalapa, Oaxaca, Mexico, sometime during Satornino's young adult life - possibly during the Mexican Revolution, around 1918.

Participants:

Germán Sánchez Morales: role: Speaker, year-born:1940; sex: M; native-language: ZOH; other-languages: SPN, ZAP

Heidi Johnson: role:Researcher, Depositor, year-born:1956; sex:F; native-language: ENG; other-languages: SPN

Example, cont.: Media details for one file

ZOH001R018I001.wav — Language: ZOH

Content type: primary_text — Archival-object: No

Access level: 1 (public)

Length: 00:09:01 — File size: 46M

Format: 44.1/24

Digi platform: PC, Flying Cow, SoundForge

Original medium: cassette tape

Recording quality: 3 (1-5, low-high)

Archived by: haj — Archive date: 2004-12-10

Description: <details about original medium or recording problems>

Metadata - Links

- OLAC
  - http://www.languagearchives.org
- IMDI
  - http://www.mpi.nl/IMDI
- Dublin Core
  - http://es.dublincore.org/
- METS
  - http://www.loc.gov/standards/mets/
- AILLA
  - http://www.ailla.utexas.org/metadata_sp.html

Citation of archive resources

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

Intellectual Property

Develop policies that address:

- liability issues for archive and host institution;
- rights of resource creators, both native speakers and researchers;
- any special conditions or concerns that creators may wish to attach to the resource;
- access and use requirements for users.

Intellectual Property

- Provide guidelines for resource producers for eliciting consent to archive & publish.
- Haga politicos con respecto a copiar y/o hacer sitio de espejo con otros archivos.

Todos las restricciones y datos sobre derechos tienen que viajar con cada recurso!

Intellectual Property - Examples

- AIATSIS:
  - http://coombs.anu.edu.au/SpecialProj/ASEDA/ASEDA.html
- AILLA:
  - http://www.ailla.utexas.org/site/use_conditions_sp.html
- OLAC:
  - http://www.language-archives.org/docs/license.html

Intellectual Property - Readings

- Lieberman article [http://www.ldc.upenn.edu/exploration/expl2000/papers/liberman/liberman.html]
- Copyright info from UT lawyer [http://www.utsystem.edu/OGC/intellectualproperty/index.htm]
- World Intellectual Property Organization [http://www.wipo.int]

Formats I

Archive may deal with 3 kinds of formats:

- Working format: what the linguist uses in the field.
- Presentation format: what we use for publication.
- **Archival format:** what we preserve.

Formats II

	Text a grammar	**Audio** a recording	**Video** a film
archival	tiff / XML	PCM wav	mpeg2
presentation	pdf / html	mp3	??
working	Excel / MS Word	minidisc	Digital camera format

Formats III

- Clearly distinguish archival formats from presentation formats so that users understand that *digital materials in presentation formats are not archive-quality materials*.
- Archives should publish their digitization standards as guidelines for producers who wish to deposit digital materials.

Formats IV

General requirements for archive-quality (master copy) formats:

- non-proprietary; that is, the encoding is in the public domain;
- portable, re-useable;
- best possible reproduction of the original.

Formats V

Archival objects (files in wav, tiff, XML, etc. formats) will be migrated forward when necessary in the future.

They'll be converted to the new format.

All the other versions will be discarded and new presentation/working formats will be generated from the new archival copies.

Audio formats I

- **PCM linear wav** is the standard audio preservation format.
- For speech, CD-quality is acceptable:
  - 44.1 kHz sample rate, 16 bit depth
- One step up: 44.1 / 24 bit depth
- For music (and for speech, according to some experts): 96 kHz / 24 bit depth

Audio formats II

(from the EMELD School:

http://emeld.org/school/classroom/audio/index.html

- Sample rate: how many times per second the sound wave is measured by the digitization software;
- Bit depth: specifies the range of numbers used to represent the sample.
- In both cases, higher values (96/24) will yield more faithful reproductions of the original.
- Higher values will also yield bigger files.

Audio formats III

File sizes for a 10-minute recording:

44.1/16	51M
44.1/24	76M
mp3 (128 kbps)	9.2M

CD-quality =~ 5M per minute

** mp3 is a compressed format

Text formats

- Goal: to be able to read the file in 10 years without searching for obsolete software.
- Experts recommend XML and Unicode (but they don't tell us how to do it yet.)
- XML & HTML are just plain text with structured tags that describe the formatting.
- Plain text and tab-separated text (output from Word, Excel, Shoebox) are perfect.

Formats - Readings

- EMELD School of Best Practice:
  - http://emeld.org/school/index.html
- For images:
- http://www.library.cornell.edu/preservation/tutorial-spanish/contents.html
- Colorado Digitization Program:
  - http://www.cdpheritage.org/resource/audio/std_audio.htm

Audio digitization equipment

- PC/Mac with 1 Gb RAM, 40+ hard disc
- Software: SoundForge (PC only), ProTools, CoolEdit
- Analog-to-digital converter: Flying Cow, Lucid, others (more expensive to connect more components at once)
- Media players
- Cables, speakers, headphones

Audio digitization procedures I

- For open-reel tapes: play them first, to determine:
  - speed, number of sides, volume levels;
  - clean and repair old splices
  - rewind tape evenly.
- Digitize a whole side and then cut the digital recording into segments (e.g. several stories on one tape) using the audio software.
- Some cleanup may be necessary to boost volume (normalize).

Audio digitization procedures II

- With SoundForge (PC), you can digitize several recordings (~10) and use the BatchConverter to produce mp3, 1-min. samples (files > 10 mins.) and cd-quality wav versions for your users.
- **It takes at least 5 hours to process 1 hour of recording.**

Archiving workflow: processing a new deposit of materials I

1. Sort the deposit into resources, assigning archive identifiers and labelling **everything** clearly.
2. Start a deposit process log to keep track of each step in the process as it's done.
3. Digitize the analog materials.
4. Do the metadata. This will probably require multiple consultations with the depositors.

Archiving workflow II

6. Produce presentation formats (mp3, pdf).
7. Upload the digital objects to the server and/or the external hard drive and/or archival CD/DVDs.
8. Link the files to the metadata.
9. Test. Best to have a different person do it.
10. Make backups daily & weekly and store them in a different location.

Databases for metadata

- mySql: http://www.mysql.org
  - + free, doc in many lgs, lots of help on the internet, not hard to learn the basics.(http://sqlzoo.net/es/)
  - - need a programmer to build the foundation (can start with AILLA's and modify that)
- FilemakerPro: $200. Lots of templates, easy to learn (relatively), no programming (exactly). PARADISEC.
- Oracle: $$$$$. Very powerful. Requires a programmer.

Archive storage

- Best is a full network server installation that someone else (professionals) manages!
- Next is a mid-range business grade server installation. Must consult locally knowledgeable people for specifics.
- Short-term (~5 years): personal computer with a 500+ Gb disc and external hard drives for backup. In 5 years you MUST transfer all files to a new drive to stay current.

Archive interfaces

- You can offer presentation copies only: cassettes, cds, texts on portable media.
- And/or make available on the Internet:
  - PHP: scripting language, free, good docs, works with mySQL, very popular worldwide
  - http://www.php.net/manual/es/
  - Commercial web-development software

How long will it take to set all this up?

- Initial website & database development: 1-2 years depending on staff, infrastructure, technology choices.
- After 2 years you will find 100 things you want to change, so plan on at least some redevelopment in 2-3 years.

Details: archive physical space

- Office space for each full-time staff position.
- Digitization lab: AILLA's lab is about 5m x 7m. Sufficient for 3 digitization stations, a worktable for sorting deposits and holding meetings, and an extra computer to use while digitizing on the others.
- Lots of shelves for storing deposits in progress.

Details: miscellaneous supplies

- CDs & cases for making copies for depositors;
- DVDs for a convenient backup format;
- external hard drive for backing up administrative computer (manager's);
- speakers, headphones – 2 each for each station;
- every kind of cable ever made;
- neat white boxes for sorting and temporarily storing deposits in progress;
- tape cleaning and splicing supplies

Details: a digitization station

Computer: Dell Dimension (PC) with 1 Gb RAM and 40 Gb hard disk.

AD converter: M-Audio Flying Cow (44.1/24)

Software: SoundForge 6.0

Media players:

- open-reel: Revox, Aiva, Uhers
- cassette: Tascam
- minidisc: Tascam

Scanner: HP

Is all this really worth it?

Absolutely yes!

A well-established archive

- Can foster “archivelets” or “jukebox archives” in local communities: an ordinary computer with a set of mp3s, refreshed from the mother archive whenever necessary;
- Can encourage comprehensive language documentation by providing a place to preserve and publish recordings;
- Can distribute educational materials over the Internet to bilingual teachers around the country.

Links

- Integración tecnológica para educación:
  - http://distancia.dgsca.unam.mx/
- “Experiencia de la red académica mexicana (CUDI)”:
  - http://lacnic.net/CUDI-MX-lacnicVI.pdf
- Centro Universitario de Investigaciones Bibliotecológicas:
  - http://cuib.unam.mx/

More links

- LINGUIST-LIST archive search engine:
- http://cf.linguistlist.org/cfdocs/new-website/LL-WorkingDirs/olac/olac-search-advanced.cfm