AILLA

The Archive of the Indigenous Languages of Latin America
The University of Texas at Austin

Information about archiving and
language documentation technologies

October 2007

Contact: Dr. Heidi Johnson
www.ailla.utexas.org
ailla@ailla.utexas.edu

Metadata = Catalog Information

Information for your project

Your name	
Your contact info	email address, snail mail address, phones, etc.
Funder(s)	
Description	a paragraph about the scope and duration of the project

Information for each item (recording, text, spreadsheet, database...)

Identifier	a unique label (see Labels page)
Title	a user-friendly descriptive title
Date of creation	use international format = YYYY-MM-DD
Place of creation	
Language	be as specific as to variety as possible
ISO code	from http://www.ethnologue.com/
Speakers' names	full names of all significant contributors
Genre keyword	e.g. narrative, wordlist, poem, oratory...
Equipment	everything you used to create this item
Relations	list all related items, e.g. interlinearized text, photographs...
Description	a paragraph about the content of the item
Length	hh:mm:ss for audio & video recordings

Information about contributors

Full name	
Nickname	e.g. initials used in transcriptions
Keep Anonymous	'Yes' if public metadata should not reveal the name
Date of birth	YYYY-MM-DD (an estimated year is sufficient)
Native language	
Other languages	other languages this contributor speaks
Place of origin	
Description	pertinent information, e.g. that the person is a curandero. Can also be used for other names, family relationships, etc.

Labels

****** Nothing could possibly be more important than labelling every single item you produce with RUTHLESS CONSISTENCY *****

(If we don't know what it is, we can't archive it.)

I. Organization criteria

Which one dominates depends on your project.

language	place (e.g. village)
date (YYYY-MM-DD)	speaker

II. Analog media (tapes, notebooks, cds, shoeboxes...)

Sort the collection e.g. by language, then by date. Number each object from 1 to N. Relate related items by means of the number (tape1 and tape2 go with notebook1.) Record these relations in your metadata catalog.

Write the labels on the objects in indelible black ink. Don't use sticky notes or colored dots for labels - they fall off.

III. Digital files

Create an extensible filename syntax for your project. Use the components to help identify place/language, speakers & researchers (initials), content type keyword (Txt), date, sequence, etc.

Folder: G_Cruz_historia

SJQ-2007_09_11-Txt_GC-acw-1.wav

SJQ-2007_09_11-Txt_GC-acw-2.wav

SJQ-2007_09_11-Txt_GC-acw.eaf

Folder: C_Flores_Cuentos

2002_08_06_Pedro-Tierra_CF1_Am.wav

IV. File system organization

A "resource" is a set of related files, e.g. recording + annotation text.

** Sort resources into separate folders to help you keep track of related items.

2-Minute Guide to Intellectual Property Rights

What are copyrights?

Copyrights govern who can use a work in what way. Copyright holders have the right to:

make copies; distribute copies;
publish; publicly display;
publicly perform; make derivative works.

Potential uses of language documentation:

- Archive with public access: easiest for speakers to access.
- Archive with restricted access:
  - password: can be shared with a small, known group.
  - time limit: restrict access until specified date (e.g. speaker's lifetime)
  - depositor control: users must contact depositor to ask for permission.
- Publish the whole work: e.g. annotated text, CD, DVD
- Publish excerpts: e.g. grammars, articles, books
- Publish derivative works: e.g. books based on texts, DVDs
- Public performance: e.g. radio broadcast
- Use whole/excerpts in classes (photocopy, download/play by professor)
- Commercial uses: books and CDs for sale

How you can document consent:

- Signed license agreements (if people are willing)
- Recorded license agreements (transcribed and translated)
- Relate agreement to resource via metadata
- Document the Codes of Conduct adhered to for future reference

Have a conversation with your consultants about uses and possible abuses; about who can/can't hear/see the materials; about possible hazards - jail, violence, embarrassment, job loss. Record the whole conversation and include it in your archive deposit.

Formats

	Archival	Presentation
Audio	wav, 44.1/16 min[1]	mp3
Video	mpeg-2 (mpg)	mov
Text: editable[2]	Unicode + txt, xml, eaf, trs, html	any
Text: non-editable	PDF/A[3]	PDF/A

Size comparison for a 10 minute recording:

5 Mb	64 kbps mp3
50 Mb	44.1/16 wav
100 Mb	mpg

1. minimum specs for speech: 44.1Khz sample rate / 16 bit depth
   for music: 96/24

2. Editable text can be downloaded from an archive and edited by users. Example: an annotation that several people contribute to. Such texts can only be preserved indefinitely if they are Unicode text that can be read by a non-proprietary program. If you don't need to allow other users to edit your texts, choose the next option.

3. PDF/A is a format standard adopted by the ISO. It's basically PDF 5.0 with no links, all fonts embedded, no animations, etc. Simple. PDF/A compliant documents can be produced with Adobe Acrobat 8.0, or you can send your texts in their original format (Word, WordPerfect, Excel, etc.) to the archive and we will convert it for you.

How to make a text that you can use forever

1. Use a Unicode font
   - Lucida Sans Unicode, Lucida Grande, Doulos SIL
2. Set up a keyboard layout that makes it easy to type your special characters:
   - SIL's Ukelele for Mac OSX:
     http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=ukelele
   - Microsoft Keyboard Layout Creator:
     http://www.microsoft.com/globaldev/tools/msklc.mspx
3. Transcribe your recordings with software that supports time-alignment, speaker turns, and tiers:
   - Transcriber: http://trans.sourceforge.net/en/presentation.php
   - Elan: http://www.mpi.nl/tools/elan.html
4. Interlinearize with a tool that outputs XML:
   - (next year's version of) SIL's Toolbox:
     http://www.sil.org/computing/toolbox/

General guidelines for creating texts:

1. Assume that someday someone else will want to look at your data and make sure that it is readable and interpretable by others.
2.Document your orthography. (The document might simply refer to the IPA.)
3. When possible, use linguistic terms and abbreviations common to your community of practice. When that's not possible, document your terms.
4. Remember that proprietary formats like Microsoft Word or Excel change almost annually. You have to be diligent about porting your files forward (opening them and saving them again) every time you upgrade your software and/or your hardware platform.
5. Save backup copies in tab- or comma-separated text format frequently.

Audio Recording Equipment

(from the School of Best Practice: http://emeld.org/school/)

1. Microphones

- Audio-Technica ATR35S ($50) headset mic
- Shure SM58 cardioid dynamic mic ($100)
- Shure Beta 53 ($600) or Sennheiser HSP2 ($450), purchased with an in-line phantom power adapter (all headsets)

2. Recorders

- M-Audio MicroTrack 24/96 ($400)
- Marantz PMD660 ($500)
- Marantz PMD 671 ($1000)
- Sound Devices 722 ($2375)

3. Other equipment

- microphone windscreen ($5-$50)
- cables, converter/connectors (e.g. XLR -> RCA)
- compact flash memory cards (~ $50 / 1 Gb)
- batteries & battery charger
- stands for microphones and lights

4. Audio digitizing/converting/editing software

- Audacity (free, but simple) http://audacity.sourceforge.net/
- Sound Forge Audio Studio ($65) PC only
- Adobe Audition ($200) Mac & PC
- Adobe Soundbooth ($200) Mac & PC

5. Video digitizing/converting/editing software

- Adobe Premiere ($65) PC
- FinalCut Express ($150) Mac

Video recording equipment

(from the School of Best Practice)

- Choose a camera that can record high-quality audio (linear PCM) at a minimum of 44.1 kHz, and be sure to set the camera to record at that sampling rate.
- If you can afford it, choose a 3-CCD camera (which has a separate CCD for each of the 3 color planes), rather than 1-CCD, because 3-CCD provides better resolution.
- Use an external microphone of good quality (the built-in mics are inadequate). Be sure to attach the microphone to the speaker's lapel, not on top of the camera.
- Always record at SP, not LP.
- Never use the digital zoom function. It's all right to use optical zoom, but be careful -- some cameras switch automatically from optical to digital zoom.
- When filming at night, it's better to switch the night shot setting to Off, because it emits enough light to make you think you're getting good video when you're actually not.
- When framing the shot, hold the camera far enough away to entirely capture all gestures made by the speaker. [But close enough so that the speaker fully occupies the screen.]
- A separate audio recorder may provide better sound quality; however, synching up the audio with the video can be difficult.
- Use the old Hollywood technique of beginning each scene with a clap (helps in synching up audio and video later) and a shot of a piece of paper showing basic metadata (date, location, name of speaker, etc.).
- A bad cable or incompatible device will be difficult to replace in the field. Test all of your equipment together *before* your trip, as well as any digitization procedures you intend to do in the field.

Recording tips and guidelines

1. If possible, set up a "recording studio" that you can use for most of your recording sessions. Protection from the wind, fabric to dampen echoes indoors, a table for your equipment, chairs, extra lights for video...
2. Worry about the wind: a pleasant breeze will turn into constant, loud, obscuring noise in your recording.
3. Unplug generators, refrigerators, and other devices that produce a low hum.
4. Practice with your equipment to learn best placement for microphones, stands, lights, chairs, etc. Especially test headset mics to make sure they're not so close to the speaker's mouth that plosives cause clipping.
5. Monitor your recordings. In a new environment, record a little, stop, listen, adjust, and then go on. And/or listen to what you recorded earlier that day and note any noisy intrusions that you might be able to change next time.
6. Make copies as soon as possible:
    - upload data from flash cards to your computer every day;
    - burn 2-3 cds for each recording. Send a copy to your archive, give a copy to your consultant, and use a copy for transcription, etc.
7. Write down the metadata (catalog information) when you make the recording.
8. Record the basic info - language, date, place, speakers' full names - at the beginning of each recording. Speak clearly in a large-scale language (e.g. Spanish). Don't forget to include your own name and the name of the language in the metadata.
9. For video, try to get some light over the speaker's head so that their face can be clearly seen.
10. Remember that video files are enormous and don't film the ground, empty chairs, etc. If there aren't any people speaking, singing or dancing, consider taking a still photo. Or, delete the uninteresting parts when you edit the video.

Digitization equipment

I. analog-to-digital converter

A. Edirol UA-5 24 bits / 96 Khz
http://www.edirol.com/products/info/ua5.html
Connects to computer with USB cable (no special sound card required.)
($240) http://www.samedaymusic.com/product--EDIUA5

II. Sound card

(if you need it for transferring digital - minidisc - input to the computer)
M-Audio Audiophile 24-Bit/96kHz
($150) http://www.bhphotovideo.com/

III. Software

1. Audacity (free & simple) http://audacity.sourceforge.net/
2. Adobe Soundbooth CS3 (PC & Mac) ($200)
http://www.adobe.com/products/audition/compare/
3. Sound Forge Audio Studio (only PC) ($60)
http://www.sonycreativesoftware.com/products/product.asp?pid=454

IV. Players

1. Cassette
Tascam 322 Dual Cassette Deck ($500)
http://www.fullcompass.com/product/301542.html

2. Minidisc (transfer to the computer through the M-Audio soundcard)
Tascam MD350 MiniDisc Recorder/Player ($470)
http://www.zzounds.com/item--TASMD350

** you need a special connector for the minidisc:
Optical to Coaxial Digital Audio Converter ($40)
http://www.cablestogo.com/product.asp?cat%5Fid=504&sku=40019

3. Reel-to-reel
You have to scrounge these wherever you can, including eBay.

V. Archival formats

1. Audio
minimum for speech: PCM WAV, 44.1Khz sample rate, 16 bit depth.
96/24 better for music.
Stereo only if the original is stereo; if it isn't, use mono.

2. Video
MPEG-2 (.mpg)

3. Texts
a. "dead" texts (that users can't edit): PDF/A
[Note: PDF/A is a new ISO standard for the pdf format. This version is bascially PDF 5.0, without links, animations, etc. It is now in the public domain, so we can guarantee that documents in this format will be readable in the future.]

b. "live" texts (that users can edit in the future):
Unicode font (Lucida Sans Unicode, Lucida Grande)
formats: txt, xml, eaf, trs, html

VI. Scanning texts
1. Canon® CanoScan® 8400F Flatbed Scanner ($200) (Office Depot)

2. Adobe Acrobat 8.0
- Professional ($130) http://www.softwaresurplus.com/Guaranteed/Low.Prices?
allows conversion of old PDFs to PDF/A.
- Standard ($120)
create new PDF/A but can't convert old ones

VII. Miscellaneous equipment
1. connectors, cables, etc.
http://www.cablestogo.com/index.asp?

2. 3.5 floppy drive ($20) http://www.tigerdirect.com/
3. CDs, labels, boxes
4. boxes for sorting deposits (we like white ones)
5. for reel-to-reel tapes: leader tape, splicing tape, splicing block
6. software for reading all sorts of digital texts: Word, WordPerfect, Excel, etc.
7. external hard drives (500 GB = $170 at Office Depot)

Where to go for more information

I. General information

- **EMELD School of Best Practice:** emeld.org/school/index.html
- **Hans Rausing Endangered Languages Project (HRELP):** www.hrelp.org/documentation/whatisit/
- **Documentation of Endangered Languages (DOBES):** www.mpi.nl/DOBES/INFOpages/applicants/dobes-ling-aspects-lang-doc.html
- **Resource Network for Linguistic Diversity:** www.linguistics.unimelb.edu.au/thieberger/RNLD.html
- **Language Archives Newsletter:** www.mpi.nl/LAN/
- **The Vermont Folklife Center:** www.vermontfolklifecenter.org/res_audioequip.htm
- **Matrix Oral History Project:** www.historicalvoices.org/oralhistory/improve-ad.html

II. Standards and organizations

Note: write to any DELAMAN member for more information

- **DELAMAN:** Digital Endangered Languages and Musics Archive Network: www.delaman.org
- **OLAC:** Open Language Archives Community: www.language-archives.org

III. Software

- **Unicode**:
  - www.unicode.org/
- **Doulos SIL Unicode IPA**:
  - www.sil.org/computing/catalog/show_software.asp?id=91
- **ELAN**, multimedia annotator:
  - www.mpi.nl/tools/
- **IMDI Editor**, java tool for IMDI metadata
  - www.mpi.nl/tools/
- **Transcriber**, transcription tool:
  - trans.sourceforge.net/en/presentation.php
- **Praat**, speech analysis tool:
  - www.fon.hum.uva.nl/praat/
- **Audacity**, free audio editing tool (can be used for digitizing):
  - audacity.sourceforge.net/
- **AILLA metadata forms** & Shoebox templates:
  - www.ailla.utexas.org/site/download_md_forms.html
- **Toolbox**, database & interlinearization:
  - www.sil.org/computing/toolbox/