Data management and archiving

University of California at Santa Barbara, June 24–27, 2008

Anthony Aristar, Helen Aristar-Dry
LINGUIST List, Eastern Michigan University

Jeff Good
University at Buffalo, The State University of New York
• Anthony Aristar, LINGUIST List, Eastern Michigan University
• Helen Aristar-Dry, LINGUIST List, Eastern Michigan University
• Jeff Good, University at Buffalo
Dual focus

- Data management to satisfy your own needs
- Data management for interoperability
  - Speaker communities
  - Other scholars
  - General public
Dual instructors

- Anthony Aristar and Helen Aristar-Dry
- Moderators of LINGUIST List
- PI’s on E-MELD, Multitree, LL-Map, LEGO
- Long-term focus on interoperability
- Jeff Good
  - Fieldwork in Cameroon
  - Long-term interest in language documentation technologies
Anthony Aristar, Helen Aristar-Dry
LINGUIST List, Eastern Michigan University
Jeff Good
University at Buffalo, The State University of New York
• **Workflow:** A model of the steps involved in carrying out a specific task

• Workflows can be devised proactively or (more often for linguists) retroactively

• Most useful for tasks which will be frequently repeated

• Assist in determination of ways to make work go more smoothly
• The “field” means different things to different people

• Many people even say they are doing “field work” even if they’ve never been to the “field”

• Your workflow will be highly dependent on your field setting
• Your office
Field settings

• Your office
• A consultant’s home
Field settings

• Your office
• A consultant’s home
• Your office
• A consultant’s home
Field settings

- Your office
- A consultant’s home
- A city/town in a less-developed country
Field settings

- Your office
- A consultant’s home
- A city/town in a less-developed country
Field settings

- Your office
- A consultant’s home
- A city/town in a less-developed country
Field settings

- Your office
- A consultant’s home
- A city/town in a less-developed country
- A village in a less-developed country
Field settings

- Your office
- A consultant’s home
- A city/town in a less-developed country
- A village in a less-developed country
- Your office
- A consultant’s home
- A city/town in a less-developed country
- A village in a less-developed country
Field settings

- Your office
- A consultant’s home
- A city/town in a less-developed country
- A village in a less-developed country
- The forest/jungle/steppe/etc.
Field settings

• Your office
• A consultant’s home
• A city/town in a less-developed country
• A village in a less-developed country
• The forest/jungle/steppe/etc.
Assumptions here

- Your field setting has sufficient access to electricity to allow you to:
  - Make more recordings than you can easily keep track of
  - Use a computer regularly

- What we’ll cover
  - Managing day-to-day workflow
  - Technological tips for the whole trip
Keep things easy

- One of the core characteristics of true field work is lack of control over
  - Your workspace
  - Your time
  - Your agenda
- Your data management workflow in the field needs to be simple and flexible
Avoid the new

- Learn new software away from the field, not in it
- Devise a workflow that capitalizes on the use of tools you already know
- Save your energy for learning the language, not a new program
- Try out all your software in a “mock” field setting (e.g., not connected to the internet; or with your laptop unplugged)
My daily workflow

• Prepare recording equipment
• Conduct sessions
• Transfer files to hard drives
• Label recordings
• Enter metadata
• Track finances
• Prepare next day’s work
• Recording techniques are the subject of other tutorials
• However, audio recording is increasingly moving to “computational” media
• Your audio equipment needs to be integrated into your digital workflow
• Recording techniques are the subject of other tutorials

• However, audio recording is increasingly moving to “computational” media

• Your audio equipment needs to be integrated into your digital workflow
Recording techniques are the subject of other tutorials.

However, audio recording is increasingly moving to “computational” media.

Your audio equipment needs to be integrated into your digital workflow.
• Memory cards are too expensive not to reuse

• You want a way to connect them to your computer using only computer power

• You need to make sure you have enough space available to handle the day’s recording

• You’ll want to be careful about file names
Conducting sessions

- Record **metadata** at the beginning of the recording
- Write down time of interesting parts recording in notebook—or clap
- Choose a file segmentation scheme that works for you
- Try to stop the recording for gossip—but remember to start it again!
• Label everything

• Digital resources: Label is the filename

• Physical resources: You’ll need labels, Sharpies, etc.

• Labels should be globally unique and informative

• The label will serve as the resource ID in metadata descriptions
Label everything

- Digital resources: Label is the filename
- Physical resources: You'll need labels, Sharpies, etc.

Labels should be globally unique and informative
The label will serve as the resource ID in metadata descriptions
Label everything

- Label everything
  - Digital resources: Label is the filename
  - Physical resources: You'll need labels, Sharpies, etc.
- Labels should be globally unique and informative
- The label will serve as the resource ID in metadata descriptions
• A good labeling scheme is
  • Informative
  • Easy to create
  • Machine-friendly
A good labeling scheme is
- Informative
- Easy to create
- Machine-friendly

Example: mff-2007-06-02-I-JCG.wav
ISOCode-YYYY-MM-DD-ID-creator.extension
A good labeling scheme is

- Informative
- Easy to create
- Machine-friendly

Labeling schemes:

- mff-2007-06-02-1-JCG.wav
- ISOCode-YYYY-MM-DD-ID-creator.extension
- Furu-2007-03-02-2-FAWG-MiniDV
- LangGroup-YYYY-MM-DD-ID-project-medium
• A good labeling scheme is
  • Informative
  • Easy to create
  • Machine-friendly

mff-2007-06-02-1-JCG.wav
ISOCode-YYYY-MM-DD-ID-creator.extension
Furu-2007-03-02-2-FAWG-MiniDV
LangGroup-YYYY-MM-DD-ID-project-medium
Munken-2005-12-1-JCG-Notebook
village-YYYY-MM-ID-creator-medium
You’ll hear more about metadata—i.e., data about your data—in the next lecture.

Different kinds of metadata:

- **Project-level metadata** (e.g., funder)
- **Session-level metadata** (e.g., date)
- **Speaker metadata** (e.g., name)
Metadata in the field

- Entering metadata can be time-consuming
- In the field, limit the scope of your metadata entry to things that change day-by-day
  - Metadata for each day’s new resources (audio, video, pictures, transcriptions, ...)
  - Speaker information
  - Resource relationships
  - Special notes/concerns
• Some kinds of relationships to encode in your metadata
  • Notebook pages and recordings
  • Multiple recordings of the same “session”
  • Pictures with speakers/sessions
  • Video recordings with audio recordings
  • Transcription with recording

Similar labels can help you keep track
Resource relationships

mff-2007-06-02-1-JCG.wav
Audio Recording

mff-2007-06-02-1-JCG.eaf
Elan Annotation File

mff-2007-06-02-1-JCG-MiniDV
Video Recording
• I don’t recommend learning a new program just for metadata

• I use a simple FileMaker database

• Excel could work well, too

• Keeping all of your metadata in one place is probably more efficient

• But an alternative would be to keep it “near” the actual files it describes
Entering metadata
Entering metadata

Menchum_Metadata

- **ID number**: Mbuk-11-6-2007-WBWG-1.wav
- **Date**: 6/11/2007
- **Speakers**: Sah Nicholas, Zang Martina
- **Content**: Grammatical elicitation of the Mbuk
- **Notes**: Second session on Mbuk in the same day,
- **Associated with**: WBWG–Notebook–2:87–97, WBWG–.wav file recorded with Marantz solid state
- **Project**: Western Beboid Working Group
- **Creator**: Jeff Good, Scott Farrar
- **Language variety**: Mbuk
• Some specialized offline metadata tools
  • IMDI Metadata Tools
    http://www.mpi.nl/IMDI/tools/
  • Field Helper
    http://acl.arts.usyd.edu.au/fieldhelper/
Specialized tools
Specialized tools
• During your field work, you don’t need to work on *all* of your metadata

• But the metadata specific to that day’s resources should be recorded each and every day (or as soon as possible)

• It may be hard to even remember what your consultant looked like—let alone what their name was—once you get back home
• Maybe the most surprising part of my workflow is “accounting”

• Much time in the field can be spent on things that seem tangential to “field work”

• Make sure you have the tools to support it
  • Spreadsheet software
  • Offline e-mail program
  • PDFs of important reference works
• You need a way to backup materials in the field

• Estimate the total amount of data you will collect

• Assess backup possibilities for your field setting
• You need a way to backup materials in the field

• Estimate the total amount of data you will collect

• Assess backup possibilities for your field setting
• You need a way to backup materials in the field

• Estimate the total amount of data you will collect

• Assess backup possibilities for your field setting
• Do not rely on one piece of hardware (e.g., a laptop) for storage

• Laptop hard drive, external hard drives, blank DVDs plus DVD burner...

• Field conditions may cause things to break faster than they would at home
• Make sure you have all the software and accessories you will need for backup

• DVD burning software, cables for video transfer, cables for external hard drives, etc.

• Redundancy—if you can spare the weight—is a good thing

• Your USB port may blow out; so take a FireWire cable—and vice-versa

• Get multiple programs for the same task
Must have tools

- **Praat**: For troubleshooting problems with audio recordings and evaluating recording quality
- **VLC**: Free, cross-platform media player (good for when things go wrong)
- Phonetic transcription tools
- Database or spreadsheet software
- Time-aligned transcription software
Gifts won’t be part of your day-to-day workflow

But the people you work with will likely want some kind of memento of your visit

Find out what technology they use to see if you can give them copies of what you collect before you leave

Spare blank CDs will probably come in handy
• Stick with what you know
• Keep things simple
• Make backup and metadata entry a habit
• Be prepared for
  • Unpredictable social environments
  • Unpredictable technological environments
  • Plan A to fail—so have a plan B (and maybe even C)
• Map out your workflow—or planned workflow—in the field
• See how software you’re already using fits into that workflow
• Determine if there’s any new kind of software you need to acquire or learn