Building a Regional Digital Language Archive for Amazonian Languages

Vilacy Galucio & Rosielle Costa & Sebastian Drude
Museu Goeldi-MCT / Belém-Brazil; Goethe-Universität Frankfurt

Brazil’s language situation
- Over 150 languages in Brazil
- 15 families and 7 isolates/unclassified
- All are under pressure, many seriously endangered
- The median language size is below 250 speakers per language

The Museu Goeldi
- Strong profile in language documentation
- Over 1000 physical objects (tapes, CDs,...)
- Since ca. 2004: digitalization
- 2008: server with Language Archiving Technology
- Part of an emerging South-American network

Capacity building
- Training in filming, recording, edition for indigenous people in the field and in the lab

Digital Language Archive (physical structure)
- 2 storage arrays with 2TB in RAID5
- 1 network storage array with 8TB in RAID5
- 1 server with 6 HD SAS with 300GB each (needs to be increased soon)
- 1 ext. tape drive LTO4 with SAS technology

Challenges & solutions
- Amount of data, storage, backups (4TB so far)
- Long time preservation and accessibility: central server with LAT is easier to update and mirror
- Getting lost in the data: organization, conventions for file hierarchy and naming
- Time and personnel constraints

Administration and metadata
We opted for the IMDI-metadata standard (XML files).
To minimize the work load for researchers and technicians we streamlined the workflow for metadata generation

APPLIED TECHNIQUES
- IMDI-Metadata (Editor, ARBIL)
- Few obligatory fields in the IMDI-template
- Individual & central templates and repositories for recurrent information (projects, speakers, languages)
- Excel tables (generated with Total Commander and Emacs Lisp-scripts)
- Emacs Lisp-scripts interact with imdi-files for, checking, cleaning, completion

Informative WEB-site on language documentation and portal to the archive (being implemented)
http://www.museu-goeldi.br/linguistica

References
More information: linguistica@museu-goeldi.br

standard for files representing entire tapes
AWE 20090526 SD-WA 01e12.wav
standard for session and resources
AWE 20090526 SD-WA 1basica400.wav
AWE 20090526 SD-WA 1basica400 animais.KWIK
ISO 639 language code
Date of recording (day and month are optional)
Sequence of tapes from the same day (only if necessary)
Code for researcher (2 or 3 digits or letters; if needed)
Code for speaker (2 or 3 digits or letters; if needed)
Sequence of tapes of one fieldtrip
Descriptive identification of session (content)
Concrete track identification (sequence # and/or description)
Type/format of file