2006 E-MELD Workshop Panel Discussion:
Standards and Tool Support for Archiving

Helping linguistic archives to better serve the documentary linguistics community

Robert E. Vann
Western Michigan University

1) What standards are documentary linguists looking for in an archive and across archives?
   - How do issues of content such as archival decisions regarding the languages and the types of language data to be archived relate to the likely users and usages of linguistic archives?
   - How do issues of access such as resource restrictions, fees for service, audience design and user friendliness of tools/linguistic archive web sites relate to the discovery/exploitation of linguistic resources?

2) What simultaneous characteristics would an ideal linguistic archive have to have to best serve the documentary linguistics community? (Examples that come close to best practice in parentheses)
   - commitment to accepting documentation regarding all human languages, endangered or not (ATM)
   - interactive organization of language resource content (LDC)
     - by language
     - by standard types of language data
       e.g., primary text audio/video, lexicons, metadata, annotations, transcriptions, etc.
     - by standard types of speech data, if any
       e.g., legacy text, word list, spontaneous conversation, linguistic interview, story, song, etc.
     - by standard types of consultants
       e.g., naturally occurring social networks, random sample, etc.
     - by depositor purposes
       e.g., endangered language documentation, sociolinguistics, dialectology, ethnography, etc.
       e.g., speech recognition, information retrieval, machine translation, etc.
   - easy online access to linguistic resources for generic users (AILLA)
     - standard restriction options
       e.g., open, restricted via web forms/passwords, restricted by permissions, closed, etc.
     - no fees ever
     - standard statements of archival audience design in the introductory pages of archive web sites and throughout archive catalogs to promote quick and easy discovery
       e.g., linguists, anthropologists, educators, language engineers, computer scientists, etc.
     - user-friendly, nontechnically oriented, nondiscipline-specific archive web site interfaces that use only nonproprietary resource formatting and that do not require any custom annotation tools, platforms, or software
       e.g., ~~CHAT~~ vs. PDF, ~~CLAN, IMDI browser~~ vs. Netscape, etc.

3) What can we do to encourage current archives to better serve the documentary linguistics community?
   - enable one-stop comparison shopping (OLAC on steroids) to catalog and search linguistic archives by language, type of language data, type of speech data, type of consultant/data producer, depositor purposes, access restrictions, fees for access, audience designs, and user friendliness
   - overt international accreditation of linguistic archives with certification of archival content and access
     - some new organization created to represent linguistic archives around the world
     - LINGUIST list through its school of best practice (win-win-win)
       - archives that wished to be endorsed by the largest linguistic organization in the world and the discipline's central electronic publication would actively accommodate their practices to fall in line with approved standards
       - linguists, as both depositors and users, would appreciate the immediate value inherent in linguistic archives accredited by the LINGUIST school of best practice
       - users outside the fields of linguistics would feel secure when visiting an archive for the first time upon seeing the official LINGUIST seal
     - consistency in terms of content and access across archives

Online archives and resources for linguistic documentation

AILLA: Archive of the Indigenous Languages of Latin America
(http://www.ailla.utexas.org/site/welcome.html)
ANLC: Alaska Native Language Center
(http://www.uaf.edu/anlc/)
ASEDA: Aboriginal Studies Electronic Data Archive
(http://www1.aiatsis.gov.au/aseda/specialproj/aseda/index.html)
ATM: Archives of Traditional Music
(http://www.indiana.edu/~libarchm/index.html)
AALF: Audio Archive of Linguistic Fieldwork
(http://www.mip.berkeley.edu/blc/la/)
CCSP: Comparative Corpus of Spoken Portuguese
(http://www.ime.usp.br/~tycho/)
CHILDES: CHIld Language Data Exchange System
(http://childes.psy.cmu.edu/)
CLAN: Child Language ANalysis
(http://childes.psy.cmu.edu)
DELAMAN: Digital Endangered Languages And Musics Archive Network
(http://www.delaman.org/)
DOBES: DOkumentation BEdrohter Sprachen
(http://www.mpi.nl/dobes/)
ELRA: European Language Resources Association
(http://www.elra.info/)
E-MELD: Electronic Metastructure for Endangered Languages Data
(http://emeld.org/)
LACITO: LAngues & CIvilisations à Tradition Orale
(http://lacito.vjf.cnrs.fr/archivage/index.html)
LDC: Linguistic Data Consortium
(http://www.ldc.upenn.edu/)
LINGUIST: The Linguist list
(http://www.linguistlist.org/)
LPCA: Language and Popular Culture in Africa
(http://www2.fmg.uva.nl/lpca/)
MPI: Max Planck Institute
(http://www.mpi.nl/world/)
OLAC: Open Language Archive Community
(http://www.language-archives.org)
OTA: Oxford Text Archive
(http://ota.ahds.ac.uk/ota/)
PARADISEC: Pacific And Regional Archive for DIgital Sources in Endangered Cultures
(http://paradisec.org.au/)
Rosetta: The Rosetta project
(http://www.rosettaproject.org/)
SAA: Speech Accent Archive
(http://accent.gmu.edu/)
Sinica: Academia Sinica Balanced Corpus of Modern Chinese
(http://www.sinica.edu.tw/ftms-bin/kiwi1/mkiwi.sh?language=1)
THDL: Tibetan and Himalayan Digital Library
(http://www.thdl.org/)
UHLCS: University of Helsinki Language Corpus Server
(http://www.ling.helsinki.fi/uhlcs/)