Charting Vanishing Voices:
A Collaborative Workshop t
Endangered Oral Cultures

World Oral Literature Project
2012 Workshop
CRASSH, Cambridge

Sustainable Solutions for Endangered Languages Data: The Language Archive

Sebastian Drude, Daan Broeder, Paul Trilsbeek
The Language Archive – Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

Topics

- Introduction
- Sustainable data from linguistic fieldwork
- The Language Archive (TLA) @ MPI-PL
- Language Archiving Technology (LAT)
- Open access, legal & ethical issues
- Summing up: key challenges for sustainable data

Introduction: TLA

- Work of the Technical Group at the Max-Planck-Institute for Psycholinguistics in Nijmegen
- Now a new unit @ MPI: "The Language Archive"
- First as institutional solution for digital data (experiments, CHILDES, ESF-SL, fieldwork)
- From 2000 on: Central archive and technical centre of the DOBES programme (documentation of endangered languages)
- Integration with other centers and European data and research infrastructures

Introduction: DOBES

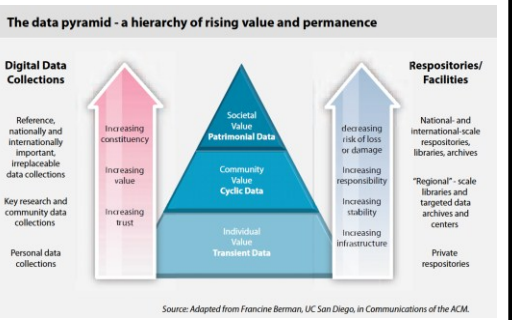
- Initiative by the VolkswagenStiftung together with German linguists
- DGFS summer school 1993, first DOBES call 1999
- Independent research teams, steering committee, advisory boards
- The heart is one central technical project and archive at the MPI Nijmegen, now "TLA"
- Total of ca. 65 individual projects (28Mio €) on about 90 target languages
- Programme will end around 2016 (>15 years)

Introduction: DOBES

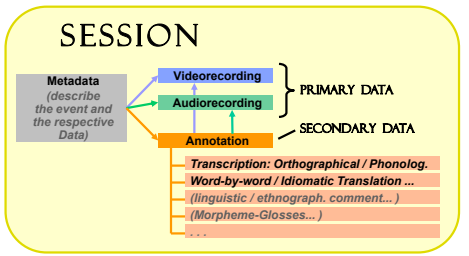
DOBES main features:

- Focus on data (linguistic analysis, revitalization and other activities welcome but additional)
- Language documentation in cultural context
- Interdisciplinary (e.g., Anthropology, Music-ethnology, Archaeology...)
- Partnership with community, training
- Emphasis on legal and ethical questions
- Common methodology and workflow
- Dissemination of language documentation (training courses, workshops, book)

Sustainable data from linguistic fieldwork



Sustainable data from linguistic fieldwork



Sustainable data from linguistic fieldwork

Challenge of ***sustainability***:

- ***Physical level***: limited lifetime of carriers
  → constant copying and replacement of carriers
- ***Logical level***: limited lifetime of formats
  → adherence to standards (Unicode, XML, open formats), constant updating of encodings
- Careful with transformations (lossy encodings, artefacts being introduced...), provenance info

Physical archives: "don't touch!"

Digital archives: "touch frequently"

The Language Archive (TLA) @MPI-PL

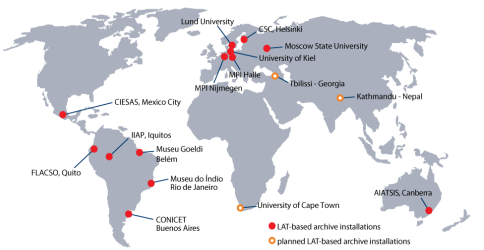
- Currently: 80TB data in well-structured sessions
- PIDs (DOI / Handles) for all resources (versions), checksum... , implementing policy rules
- Data on ca. 200 languages, & CHILDES, Dutch...
- DOBES: 25TB on ca. 60 languages
- All data is online accessible (with access rights)
- Software and infrastructure development depends on project funding
- Establishing "The Language Archive" aims at a long-term perspective for a sustainable archive

The Language Archive (TLA) @MPI-PL

DOBES field sites

The Language Archive (TLA) @MPI-PL

"Regional" LAT archives

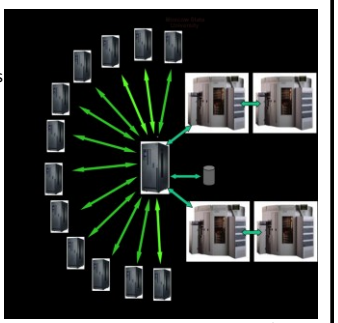


The Language Archive (TLA) @MPI-PL

Growing number of archives using LAT

Six automatic full copies at three locations in Germany

Institutional guarantee for bitstream-preser-vation by the MPG for 50 years



The Language Archive (TLA) @MPI-PL

- Primary data: uncompressed PCM audio, MPEG video, in future jMPEG2000 (lossless compressed)
- Secondary data: Elan Annotation Format (XML-based, Unicode), “standard format” (Toolbox), and other open formats, also PDF
- Metadata: IMDI standard (in future: CMDI)
- Based on an integrated set of tools for archive administration and access, the “Language Archiving Technology” (LAT) suite of tools
- Regional archives based on LAT are being set up

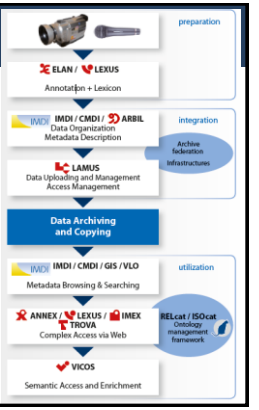
The Language Archive (TLA) @MPI-PL

Collaboration in larger projects (applying LAT):

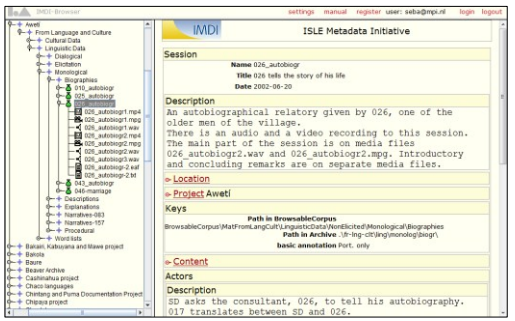
- Leading role in different EU projects working on developing e-science infrastructure for the humanities (digital humanities / “eHumanities”)
- CLARIN (Common Language and Technology Research Infrastructure): Lang. Res. & Techn.
- DASISH (Data Service Infrastructure for the Social Sciences and Humanities)
- European Strategy Forum on Research Infrastructures (ESFRI)
- Cooperation outside Europe (DELAMAN, RELISH)

Language Archiving Technology (LAT)

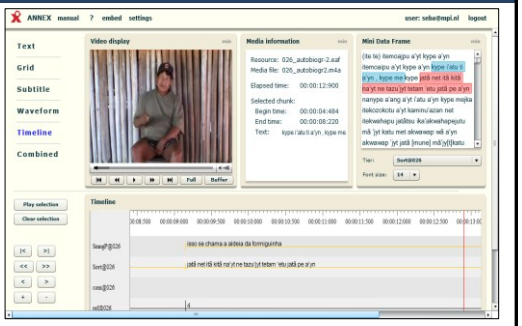
tool	state
LAMUS	mature
AMS	mature
IMDI	oldish
ARBIL	mature
CMDI	in progress
ELAN	mature
ANNEX	mature
IMEX	mature
TROVA	mature
LEXUS	redesign
VICOS	redesign
ISOcat	mature
Bridge	started
ADDIT	taken out



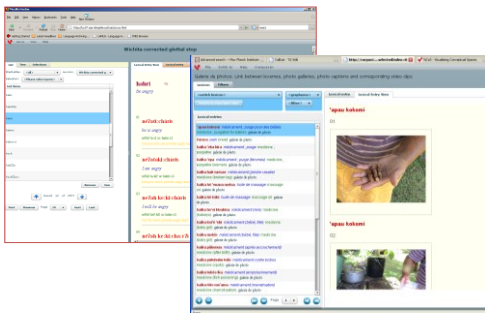
Language Archiving Technology (LAT)



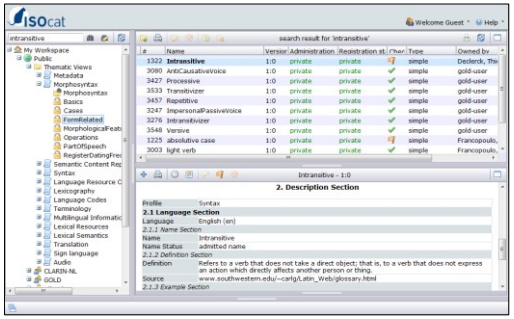
Language Archiving Technology (LAT)



Language Archiving Technology (LAT)



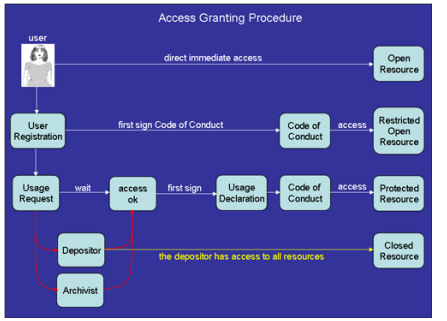
Language Archiving Technology (LAT)



Open access, legal & ethical issues

- Open access to research results ***and data***
  (Berlin declaration on Open Access to Scientific Knowledge)
- Accountability: EL data are irreproducible
- But: respect for privacy of human subjects
- Informed consent and anonymisation?
- Legal situation is complicated for all online-resources (national vs. international law etc.)
- DOBES: legal and ethical considerations are important (code of conduct, agreements, LAB)
- Trust between all parties is of key importance

Open access, legal & ethical issues



Summing up: sustainable data

- Longevity:
  (1) bit-stream → copies, migration,
  (2) interpretability → standards, format update
- Access:
  (1) identify & locate → metadata, search tools,
  (2) retrieve & visualize content → access tools, download
- Public access: trust, Code of Conduct, responsibility, access management
- Provide data to the people we record: support for dedicated portals & enriched publications

Summing up: sustainable data

- Maximum advantage of the access to data:
  A language archive is just one component in a digital research environment, interoperability
- Embed our analyses in accessible data: planned authoring environment for scholarly work
- Standards-conformant formats: good and useful tools will attract users (ARBIL, ELAN, in future with semi-automatic interlinearization function, possibly integrated with LEXUS, LMF, ISOcat)
- Support for most stages in the lifecycle of language documentation data