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
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)
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)
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
Sustainable data from linguistic fieldwork

The data pyramid - a hierarchy of rising value and permanence

Digital Data Collections
- Reference, nationally and internationally important, irreplaceable data collections
- Key research and community data collections
- Personal data collections
- Increasing constituency
- Increasing value
- Increasing trust

Societal Value Patrimonial Data
- Decreasing risk of loss or damage
- Increasing responsibility
- Increasing stability
- Increasing infrastructure

Community Value Cyclic Data
- "Regional" - scale libraries and targeted data archives and centers
- Increasing infrastructure

Individual Value Transient Data
- Private repositories

Source: Adapted from Francine Berman, UC San Diego, in Communications of the ACM.
Sustainable data from linguistic fieldwork

**SESSION**

- **Metadata**
  - (describe the event and the respective Data)

- **Videorecording**
  - PRIMARY DATA

- **Audiorecording**
  - PRIMARY DATA

- **Annotation**
  - SECONDARY DATA

- **Transcription: Orthographical / Phonological**
- **Word-by-word / Idiomatic Translation ...**
  - (linguistic / ethnographic comment... )
  - (Morpheme-Glosses... )
  - ...

Drude, Broeder, Trilsbeek

The Language Archive
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”
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
• 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
“Regional” LAT archives

- Lund University
- CSC, Helsinki
- Moscow State University
- University of Kiel
- MPI Halle
- Tbilissi - Georgia
- Kathmandu - Nepal
- FLACSO, Quito
- IIAP, Iquitos
- Museu Goeldi, Belém
- Museu do Índio, Rio de Janeiro
- University of Cape Town
- AIATSIS, Canberra
- CONICET, Buenos Aires

- LAT-based archive installations
- planned LAT-based archive installations
Growing number of archives using LAT

Six automatic full copies at three locations in Germany

Institutional guarantee for bitstream-preservation by the MPG for 50 years
• 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
Collaboration in larger projects (applying LAT): 

- Leading role in different EU projects working on developing e-science infrastructure for the humanities (digital humanities / “eHumanities”) 


- DASISH (Data Service Infrastructure for the Social Sciences and Humanities) 

- European Strategy Forum on Research Infrastructures (ESFRI) 

- Cooperation outside Europe (DELAMAN, RELISH)
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
<table>
<thead>
<tr>
<th>tool</th>
<th>state</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAMUS</td>
<td>mature</td>
</tr>
<tr>
<td>AMS</td>
<td>mature</td>
</tr>
<tr>
<td>IMDI</td>
<td>oldish</td>
</tr>
<tr>
<td>ARBIL</td>
<td>mature</td>
</tr>
<tr>
<td>CMDI</td>
<td>in progress</td>
</tr>
<tr>
<td>ELAN</td>
<td>mature</td>
</tr>
<tr>
<td>ANNEX</td>
<td>mature</td>
</tr>
<tr>
<td>IMEX</td>
<td>mature</td>
</tr>
<tr>
<td>TROVA</td>
<td>mature</td>
</tr>
<tr>
<td>LEXUS</td>
<td>redesign</td>
</tr>
<tr>
<td>VICOS</td>
<td>redesign</td>
</tr>
<tr>
<td>ISOcat</td>
<td>mature</td>
</tr>
<tr>
<td>Bridge</td>
<td>started</td>
</tr>
<tr>
<td>ADDIT</td>
<td>taken out</td>
</tr>
</tbody>
</table>
An autobiographical relatory given by 026, one of the older men of the village. There is an audio and a video recording to this session. The main part of the session is on media files 026_autobiogr2.wav and 026_autobiogr2.mpg. Introductory and concluding remarks are on separate media files.

**Location**

**Project** Aweti

**Keys**

Path in BrowsableCorpus
BrowsableCorpus\MatFromLangCult\LinguisticData\NonElicited\Monological\Biographies
Path in Archive .\fr-ing-clt\ling\monolog\biogr

**Content**

**Actors**

**Description**

SD asks the consultant, 026, to tell his autobiography. 017 translates between SD and 026.
Language Archiving Technology (LAT)

Drude, Broeder, Trilsbeek

The Language Archive

CRASSH, 2012/06, No. 29
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
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
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**
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

Drude, Broeder, Trilsbeek
The Language Archive
CRASSH, 2012/06, No. 34
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
Charting Vanishing Voices: A Collaborative Workshop to Map 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