DOBES/MPI Archive Issues

Peter Wittenburg

MPI for Psycholinguistics

DoBES Archive
(DoKumentation BEdrohter Sprachen)
(Documentation of Endangered Languages)
(funded by the VolkswagenFoundation)
some DOBES Numbers

• from 7 years of DOBES we learned a lot
  • teams worked very hard
  • recording, eliciting, annotating, managing, etc data is a hard job

• average nr. of session hours / team: audio (59 h), video (72 h),
• transcription average: 1:35, translation average: 1:25
• deep analysis about 1:100 (morphosyntax, not even gesture or so)
• 131 h fully: 2600 working days = 8 years (if not crazy beforehand)

• average: transcription (50h), translation (29h), deep (14h)
• ~ about 484 working days only for the linguistic analysis

• this all means that there is still so much to do for PhDs, students, ...
• this also means that the archive needs to be open for all sorts of extensions
The questions

• What are the bottlenecks in creating archives?

• Are archives used to the extent imagined for them or are they underutilized?

• What could increase their effectiveness?

First little information – very short (for more see flyer)
State, Preservation & Distribution

- at MPI about 30 Terabyte
- >250,000 resources
- ~30,000 hours recording
- 60 Mio annotations
- DOBES about 10%

- all 5 years new technology
- 4 copies at large centers

- synchronized regional archives are essential
  - another copy
  - data back to regions
  - trust

MPI
Leipzig

2 Computer Centers in Munich (one from MPG)

2 Computer Centers in Göttingen (one from MPG)

IIAP
Iquitos

CONICET
BA

Mdl
Rio

University
Lund

University
Kiel

SOAS
London

AIATSIS
Canberra

MPI Nijmegen

2 Copies

Belem, Tbilisi, Timor, Bangkok, Windhoek, Katmandu, Birmingham, Berlin, Halle, ...

CONICET
BA

AIATSIS
Canberra

MPI Nijmegen

2 Computer Centers in Munich (one from MPG)

2 Computer Centers in Göttingen (one from MPG)
Many Different Contributors

- 40 language documentation teams from the DOBES program
- about 50 MPI field researchers + child language + SL/gesture + ...
- increasingly more external researchers due to MPI service
LAT Dimensions: Life Cycle Support

Shoebox
CHAT
Transcriber
some XML?

many smart developers

ELAN/LEXUS/SYNPATHY
Annotation + Lexicon

IMDI
Data Organization, Metadata

LAMUS
Data Uploading and Management
Access Management

Data Archiving and Copying

IMDI / GIS
Metadata Browsing & Searching

ANNEX/LEXUS/SEARCH
Complex Access via Web

ADDIT/VICOS/MEL
Enrichments/Views

ODIT/ISOcat
Ontology management framework

online changes via the web

preparation
integration
offline changes
utilization

Peter Wittenberg
DOBES
NSF Documenting Endangered Languages Workshop,
Durham, New Hampshire, October 2007
Accessibility

Legal & Ethical Issues

- very sensitive issue and many different entities/persons are involved
- rules are not transparent - much depends on trust
- researcher/depositor is focal point for all access matters
- clarify attitude of archive (just right to archive, no copyright)
- code of conduct as basis of behavior for all

• still too much is closed
• why is it closed?
• need usage projects
• **virtual collections** across archives as goal
• all to be based on appropriate agreements to establish trust
• AAI technology tested and ready
Question 1

- What are the bottlenecks in creating (&maintaining) archives?

- what is an archive?  traditional or Live Archive

- technologists
  - can we process old formats – still have many old recorders
  - need a robust “machinery” – ours is (getting) ready
  - need a bitstream survival strategy (migration, distribution)
  - need a strategy for interpretability (-> standards, conversion)
  - need continuous funding (HW,SW)
  - need a good and professional team over years

- researchers
  - researchers start understanding the message “to give copies”
    UNESCO: 80% of recordings are endangered
  - are researchers willing to spent time on MD creation etc
    basically: invest time for others
  - for most of researchers still all is very abstract
  - some difficult questions with no answers remain
Question 2

• Are archives used to the extent imagined for them or are they underutilized?

• clear answer: they are not used yet as they could
• again which concept: traditional or Live Archive

• browsing/searching in metadata is boring – but MD is necessary
• still live in the “download phase”
  • then only advantage is well-organized domain – no chaos
• is there an added value of digital archives?
• yes if researchers (and others)
  • link ePublications with resource fragments (done)
  • use the geographic paradigm on a shared level
  • can enrich holding in many dimensions
  • in particular: commentary and relations
  • can create different views on material
  • create virtual collections across boundaries (even institutional)
  • etc etc
In Addition: Other “Typical” Views

- collaborative knowledge space with culturally relevant concepts for semantic navigation where concepts are center points for all sorts of information
- genealogy view to come next
Special Community Portals & GIS

• fostering the creation of special web-sites by REST interfaces and templates
• fostering the GIS presentation by special converters
many diverging interests

Technical Accessibility

- problem is that there are so many different views and interests
- can’t satisfy all expectations

<table>
<thead>
<tr>
<th></th>
<th>download</th>
<th>management</th>
<th>discovery</th>
<th>consistency</th>
<th>statistics</th>
<th>visualization</th>
<th>vcollections</th>
<th>extensions</th>
<th>permissions</th>
<th>inspections</th>
</tr>
</thead>
<tbody>
<tr>
<td>archivist</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>researcher</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>communities</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>journalists</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>funders</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>students</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>who else</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

special editions, CDROMs etc special skills are required

Live Archives

yes

Peter Wittenberg
DOBES
NSF Documenting Endangered Languages Workshop,
Durham, New Hampshire, October 2007
Question 3

• What could increase their effectiveness?
  
• What's that: cost effectiveness – attractiveness for users?
• Let's not forget: “all digital domain just started about 10 years ago”

• Cost effectiveness is important – machinery is functioning
• Training, training, training, ...
  • At MPI certain things work but not outside
  • Start in January with a course about “advanced methods”
  • But who is paying the researchers?
• Change of minds (openness, ethics, trust, ...)
• Bandwidth for online work – parallel to download culture
• Funds for **Live** Archives developments
• Allow building virtual collections -> AAI infrastructure
• Easy interoperability frameworks (GOLD, ISOcat, ODIT, ...)