

National Science Foundation  
Workshop for  
**Documenting Endangered  
Languages**



October 14 - 17, 2007  
The New England Center,  
University of New Hampshire,  
Durham, NH

All materials contained herein are for viewing only.  
Reproduction or dissemination without permission is strictly prohibited.  
For permission or more information, write to [DEL@NSF.GOV](mailto:DEL@NSF.GOV)



## 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

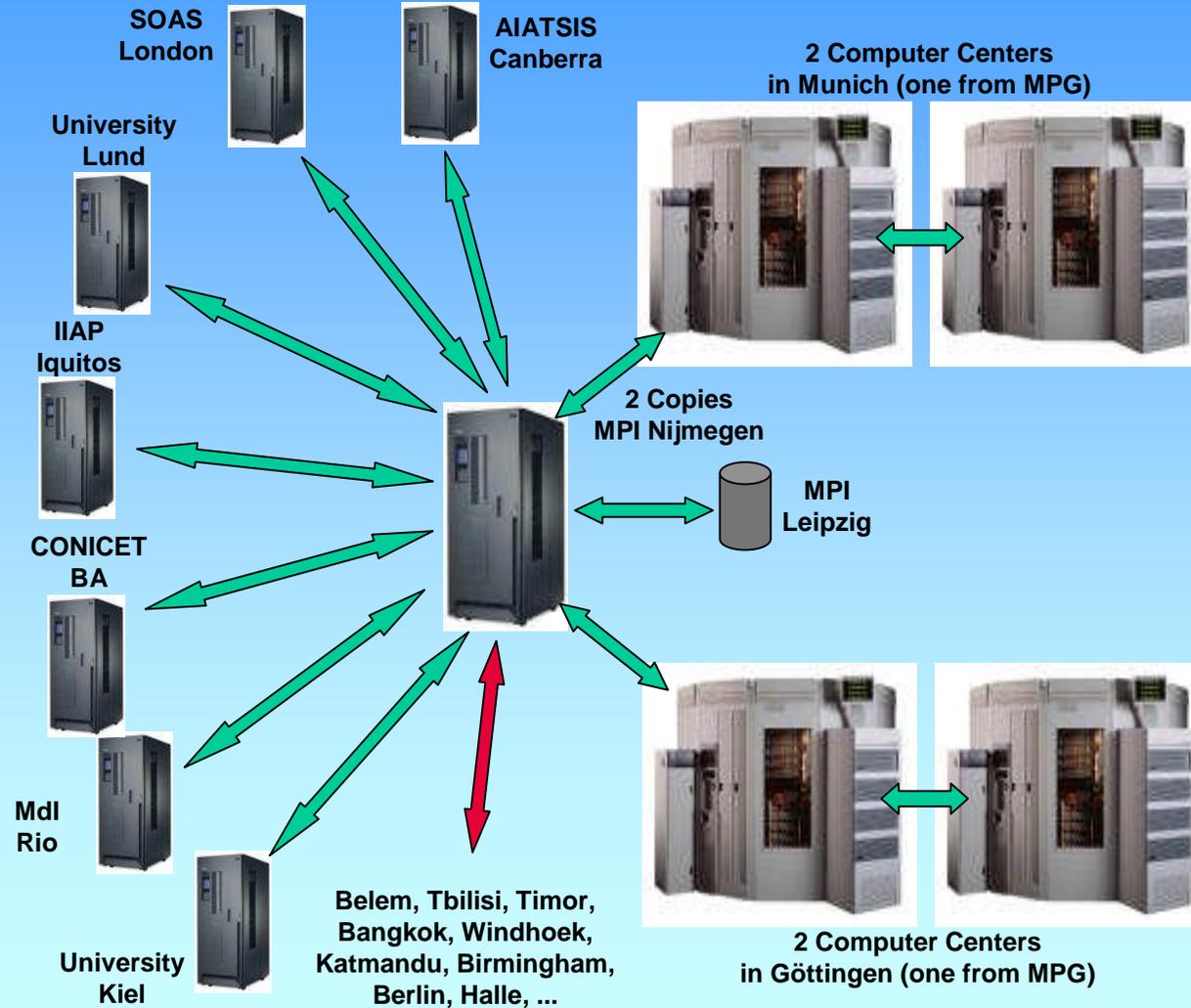


- What are the bottlenecks in creating archives?
- Are archives used to the extent imagined for them or are they underutilized?
- What could increase their effectiveness?

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



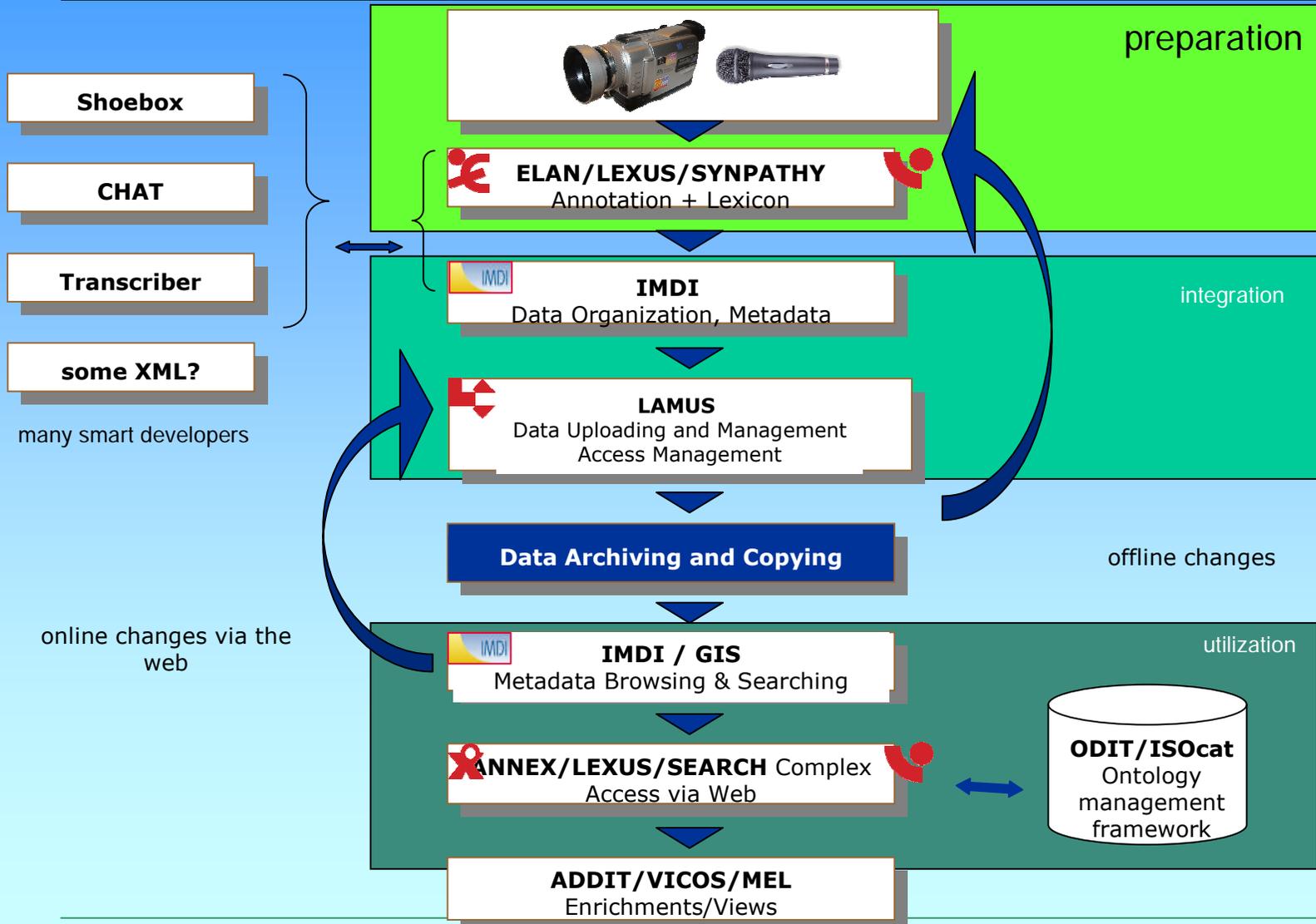
## Many Different Contributors



- 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



Peter Wittenberg  
DOBES

NSF Documenting Endangered Languages Workshop,  
Durham, New Hampshire, October 2007

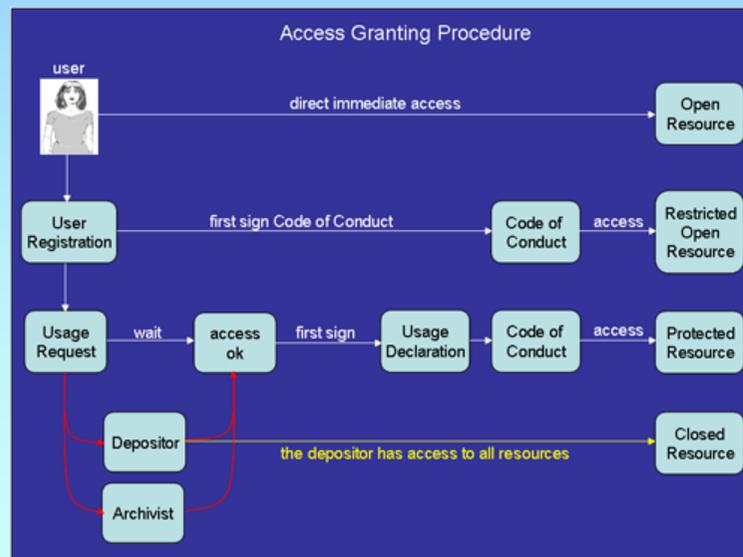


# Accessibility



## Legal & Ethical Issues

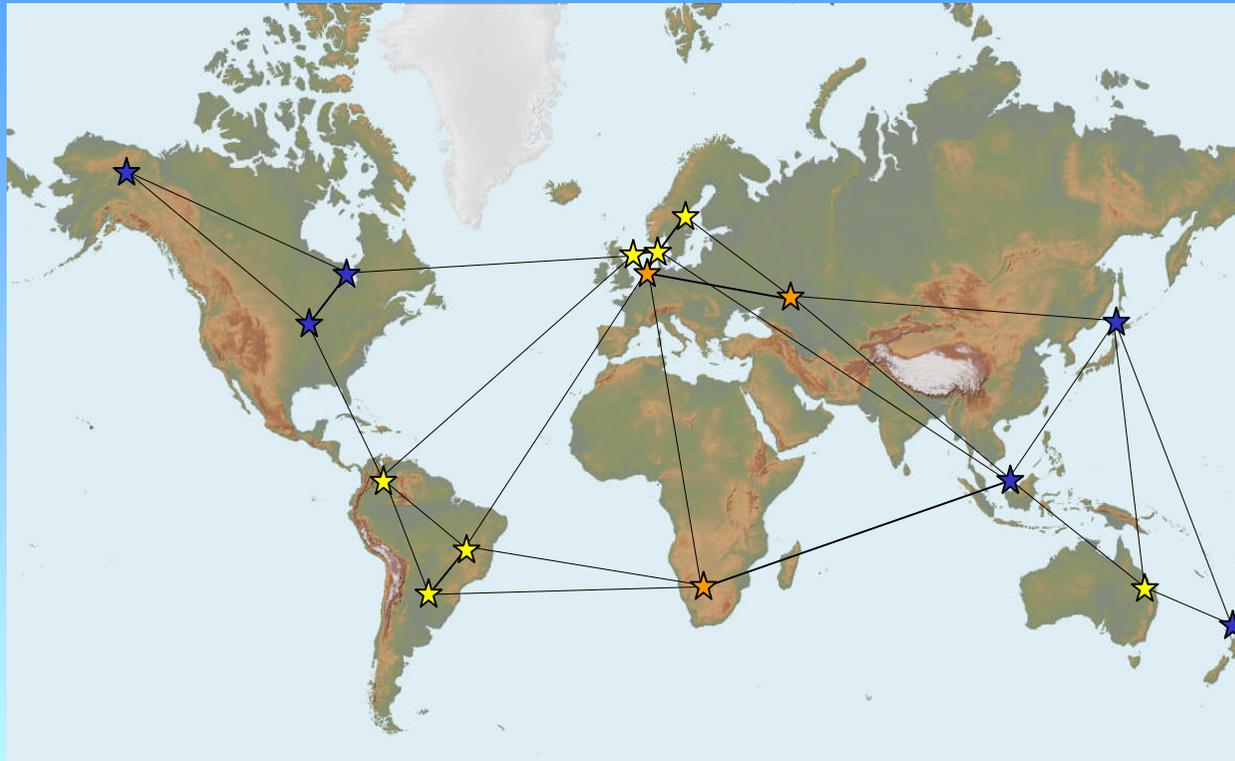
- very sensitive issue and many different entities/persons are involved
- rules are not transparent - much depends on trust
- researcher/depositor is focal point for all access matters
- clarify attitude of archive (just right to archive, no copyright)
- code of conduct as basis of behavior for all



- still too much is closed
- why is it closed?
- need usage projects



# Archive Federations



- **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

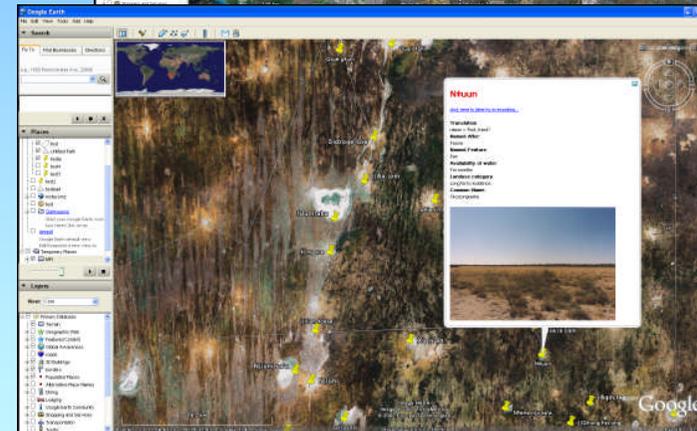
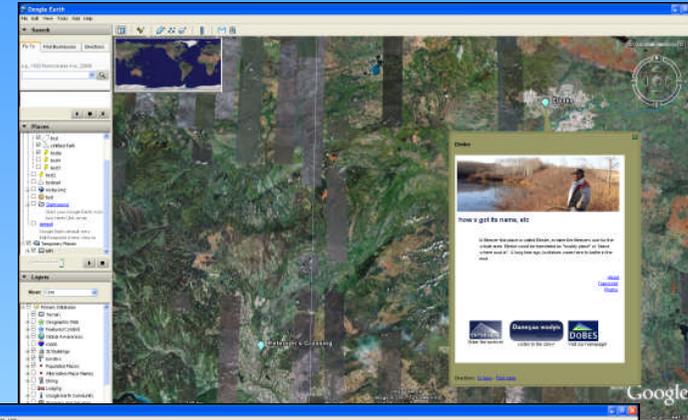
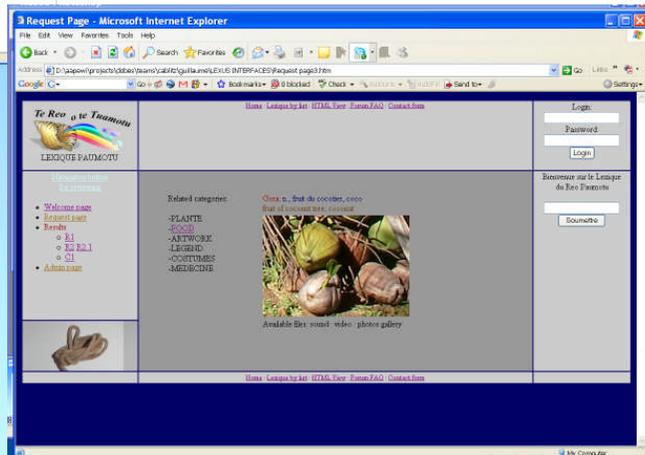


The screenshot displays the LEXUS Knowledge Spacer interface in Mozilla Firefox. The main workspace shows a knowledge space construction area with a central node 'GORA' and branches for 'FOOD', 'ARTWORK', and 'MEDICINE'. A list of noun words is on the left, and a multimedia view area at the bottom shows a video player. A lexical entry for 'ka'apa' is shown on the right, with a definition and a photograph of a cowrie shell. A video player shows a person working with a tool, and a timeline is visible below it.

- 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

Peter Wittenberg  
DOBES

NSF Documenting Endangered Languages Workshop,  
Durham, New Hampshire, October 2007



# many diverging interests



## Technical Accessibility

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

	down-load	manage-ment	dis-covery	consis-tency	statis-tics	visua-lization	vcollec-tions	exten-sions	permis-sions	inspec-tions
archivist		X	X	X	X					
researcher	X		X			X	X	X	X	
communities	X					X		X	X	
journalists			X							X
funders										X
students			X				X	X		X
who else										

special editions, CDROMs etc  
special skills are required

Live Archives  
yes



## 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, ...)