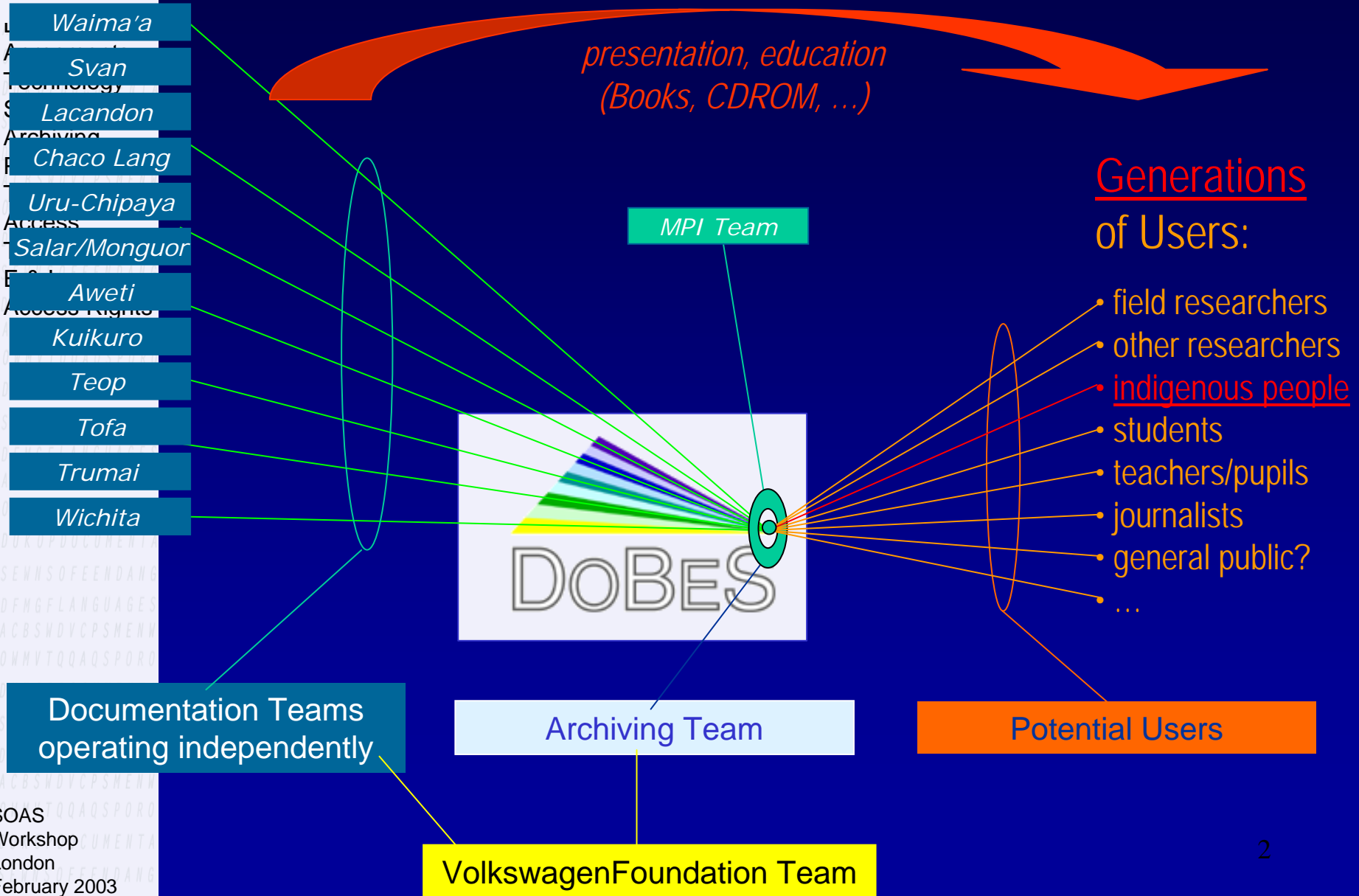


# The DOBES Model of Language Documentation

Peter Wittenburg  
Max-Planck-Institute for Psycholinguistics

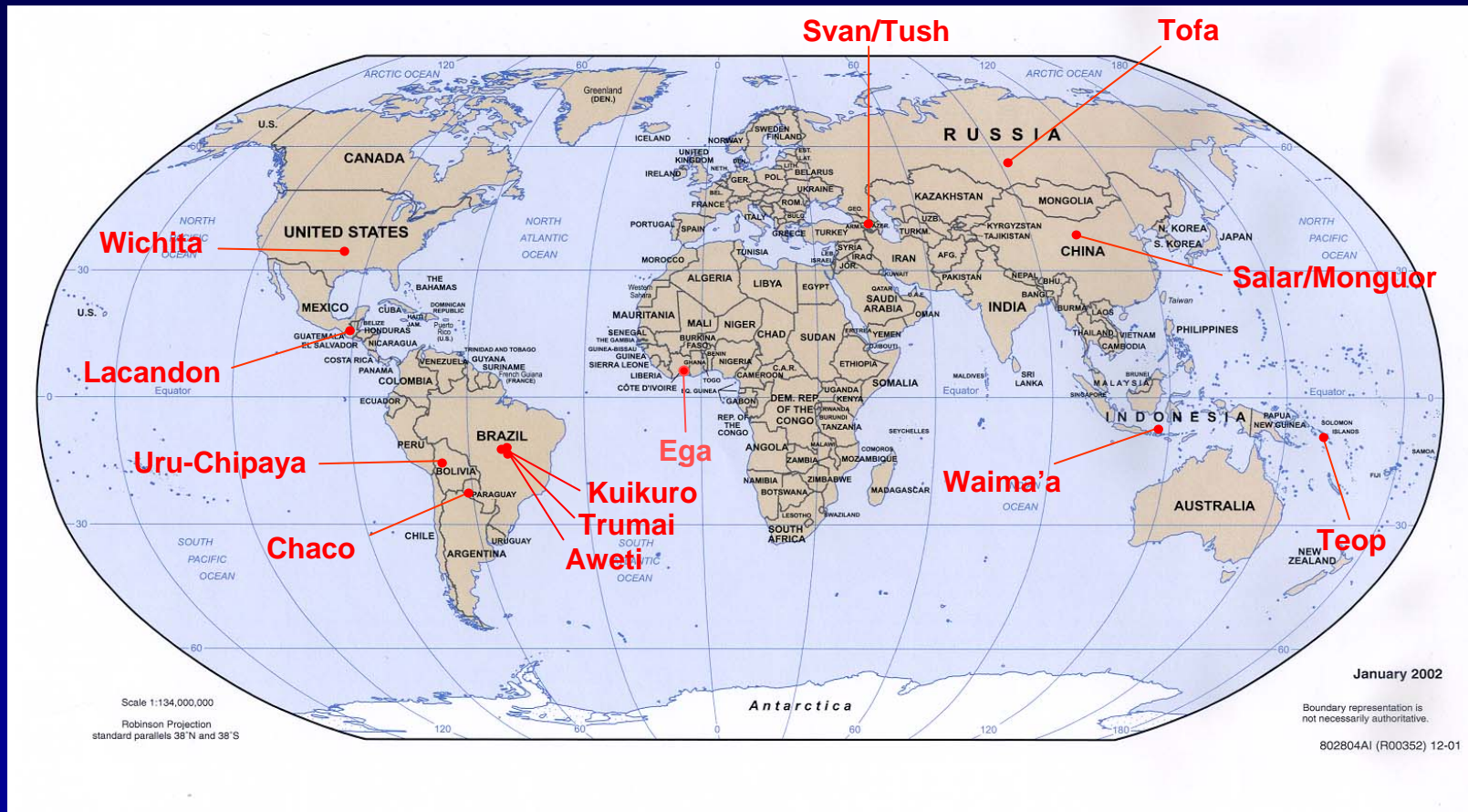
[peter.wittenburg@mpi.nl](mailto:peter.wittenburg@mpi.nl)  
[www.mpi.nl/DOBES](http://www.mpi.nl/DOBES)

# Scenario



# Languages in DOBES

Scenario  
 Languages  
 Agreements  
 Technology  
 State Archive  
 Archiving  
 Preservation  
 Tools  
 Access  
 Training  
 E & L  
 Access Rights



- started September 2000 with 8 teams in a pilot phase
- currently DOBES includes 12 teams
- expect up to 20 teams operating in parallel for the next 4 years<sup>3</sup>

# Documentation Agreements

Scenario  
Languages  
Agreements  
Technology  
State Archive  
Archiving  
Preservation  
Tools  
Access  
Training  
E & L  
Access Rights

- language documentation to preserve human heritage
- intention to cover great typological difference in DOBES
- revitalization is recognized by all participants as another major goal
- language documentation capturing also its cultural background, facial expressions, lip movements, etc
- therefore focus on multimedia recordings and inclusion of ethnologists, musicologists, ...
- a few principles / agreements (bottom up approach)
  - teams have to give (copies of) their data to the archive
  - variety of text types and genres
  - 2 main tiers (orthographic/phonetic, translation in major language)
  - further glossing in local lingua franca and English
  - for some material deep linguistic analysis (could be Adv.Glossing)
  - good documentation of all tag sets (morphosyntactic terms)
  - provision of a lexicon (not “a” to “x”, but for example topic oriented)
  - sketch grammar, field notes, ...

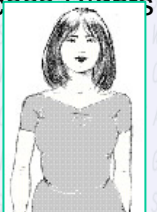
# Technical Agreements

Scenario  
 Languages  
**Agreements**  
 Technology  
 State Archive  
 Archiving  
 Preservation  
 Tools  
 Access  
 Training  
 E & L  
 Access Rights

- agreement on a workflow per team
  - utterly important: both become aware of each others special needs
  - have to take care that no resource is lost
  - have to take care each resource can be integrated
- help for archivist in organizing the material by metadata
- for the **archive** holding adherence to some standard formats
  - such as XML, UNICODE, PCM 20kHz, MPEG2, MPEG1/4, JPEG
  - not MD and DV
- for the **researcher** a recommendation to use tools that support the formats, but also support for
  - SHOEBOX nothing better yet
  - Transcriber an excellent tool for audio annotation
  - Praat an excellent tool for sound analysis
  - WORD have built a converter
  - etc

# State of Archive

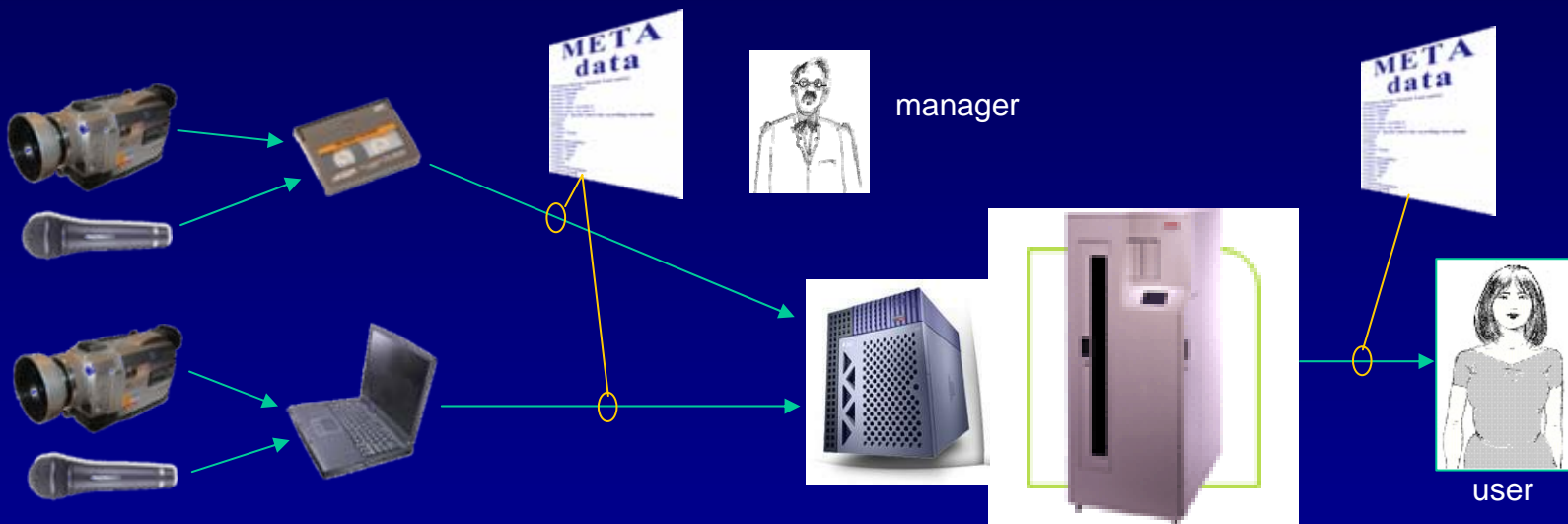
Scenario  
 Languages  
 Agreements  
 Technology  
**State Archive**  
 Archiving  
 Preservation  
 Tools  
 Access  
 Training  
 E & L  
 Access Rights



researcher

Workflow utterly important - otherwise chaos

- central: all digitization by MPI, good - but slow cycle time
- decentral: digitization in the field, error prone



350 Digital Master Files (digital copies of tapes > 350 h)  
 561 Sessions described by metadata and (therefore) organized  
 49 Annotations

many Photos + field notes  
 some first grammars and lexicons  
 many conversions from "odd" formats

MPI total > 3000 h a/v  
 > 15.000 sessions

## What are the big challenges if we agree that we need to preserve?

- shorter media lifetime vs. long-term storage  
 cuneiforms (clay tablets) -> paper -> magnetic media
  - **solution: continuous migration**
- danger of technological failures and other emergency cases
  - **solution: automatically copying of data**
  - DOBES: 2 copies at MPI, one in Leipzig , (one copy at U Nijmegen)
  - **need more - need a DATA GRID**
- guarantee availability of data independent of technology change
  - there is no good solution - but hope
  - **need standards, education, awareness + understanding, discipline**
  - **format and encoding migration**

011001010100001010110100101010 →



## The “Immediate” Way

- invest **now** more time and money
  - use experts **now** to carefully document steps and categorize etc
  - transfer to standard and well-documented formats **now**
  - train people as much as possible to follow “modern” methods
  - advantage: money effort now and well-organized archive
  - disadvantage: needs more effort and can’t take all perhaps

## The “Later” Way

- take all and invest some time and money **later**
  - take all and put it in a repository as it is
  - carefully categorize **later**
  - transfer to standard and well-documented formats **later**
  - advantage: can take all
  - disadvantage: needs money later, experts are not available

Reality:

Some way in between

In DOBES:

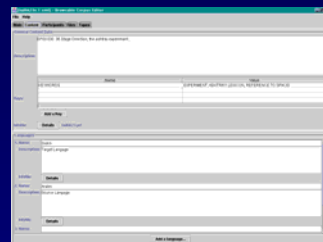
Attempt to categorize and transfer everything  
(XML, UNICODE, MPEGx, ...)

## Metadata Categorization

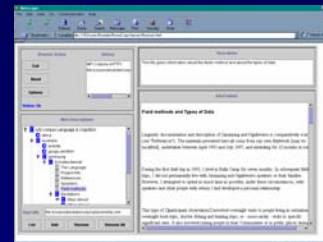
- archive needs one standard, IMDI largely influenced by DOBES
- all open and interoperable with DublinCore/OLAC



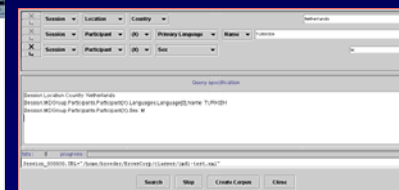
[www.mpi.nl/ISLE](http://www.mpi.nl/ISLE)



MD Editor



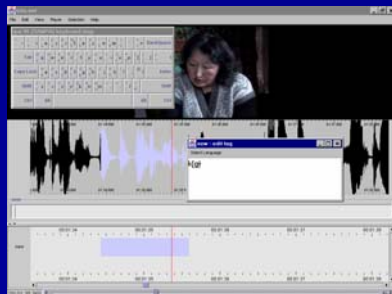
MD Browser



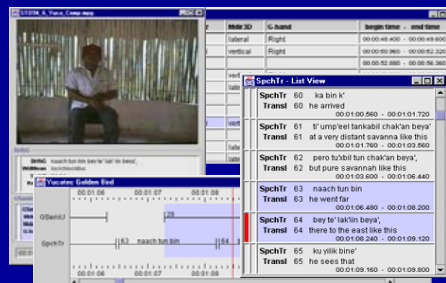
MD Search

## Annotation, Exploitation, Lexicon, Cutting, ...

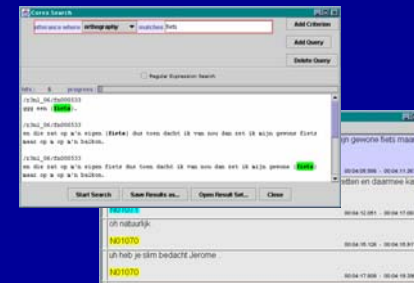
- a number of tools are recommended for different purposes
- Shoebox, Transcriber, Praat, Adobe Premiere, CoolEdit, Tsunami, ...
- accept even WORD files (special WORD to XML converter)
- develop own Multimedia Annotation Tool that people can use (or not)
- will come new ones that are interesting



ELAN Annotation



ELAN Viewers



ELAN Searching

By the way:

all MPI tools are and will be freely available for academic use

in principle all MPI tools will be made Open Source

[www.mpi.nl/tools](http://www.mpi.nl/tools)

## Two aspects for accessing archive material

- access rights issue (later)
- technical form of access

## Archivist offers shells to

- browse and search through the archive
- to exploit (and annotate) the resources

## but

- all data is addressable via Web-mechanisms (URL, HTTP)
- i.e. everyone can use his/her own ways to use the data (given suitable access rights)
- no one is bound to using the offered shells

Scenario  
Languages  
Agreements  
Technology  
State Archive  
Archiving  
Preservation  
Tools  
Access  
Training  
E & L  
Access Rights

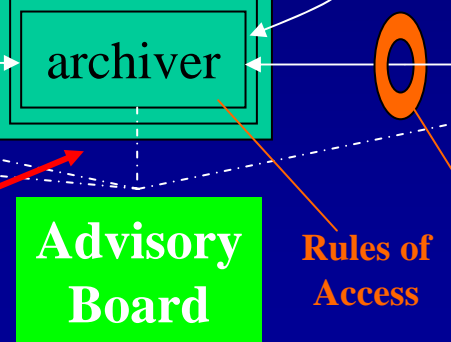
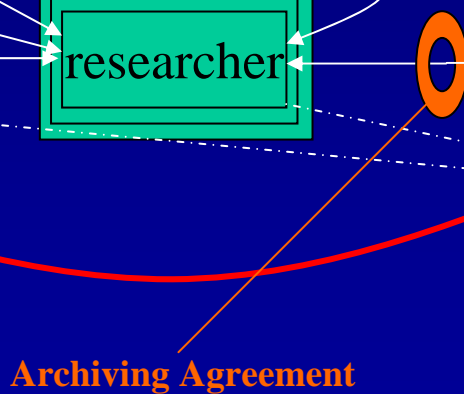
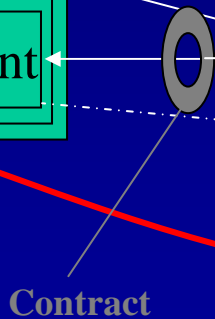
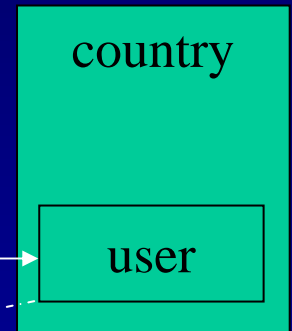
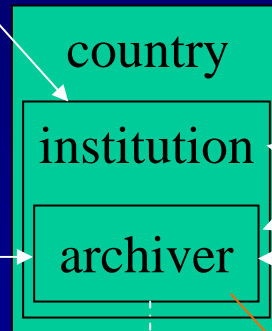
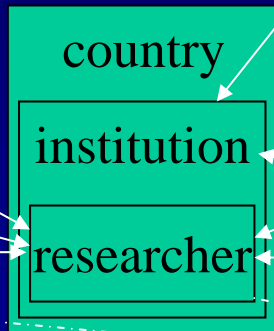
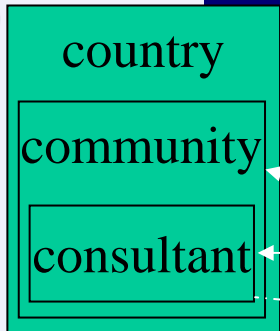
- much attention to train teams and interact with them
- we have to learn and to adapt continuously
- different skills in teams
- funny in DOBES: less tech skilled teams have high “production rate”
- focus is on
  1. learn the needs and circumstances of the teams
  2. transfer the linguistic agreements
  3. agree on workflow aspects which is “chaos” prevention
  4. transfer knowledge about practical (field work) issues
  5. training for metadata corpus organization
  6. training on digitization, annotation, exploitation, lexicon tools

# Ethical & Legal Issues

Scenario  
 Languages  
 Agreements  
 Technology  
 State Archive  
 Archiving  
 Preservation  
 Tools  
 Access  
 Training  
**E & L**  
 Access Rights



was seen as very important point  
 deserved much attention  
 ongoing discussion



## The responsible researcher is central in DOBES

- knows the community and individuals
- large differences between communities (and researchers)
- knows “our” industrialized world
- knows the archivist and most probable users
- therefore, researcher determines access rights

## However agreement of making resources as open as possible

- difference between sound/video and linguistic add-on
- temporary protection of linguistic data in case of dissertations

- metadata is open !!!

[www.mpi.nl/corpus](http://www.mpi.nl/corpus)  
[www.mpi.nl/DOBES](http://www.mpi.nl/DOBES)

May come conflict situations:

Advisory Board

Paper at the LREC Conference

# DOBES Archive

## Purpose and Implementation

Hennie Brugman

Stephen Levinson

Romuald Skiba

Peter Wittenburg

[hennie.brugman@mpi.nl](mailto:hennie.brugman@mpi.nl)



Max Planck Institute for Psycholinguistics



# Introduction

- experience and discussions from DOBES pilot phase
- some discussions also at the MPI
- DOBES pilot phase started in September 2000
- 8 documentation teams and one archiving team
- main phase started in April 2002
- 12 documentation teams and one archiving team
  
- views of the DOBES archivist
- mostly in agreement with discussions within DOBES

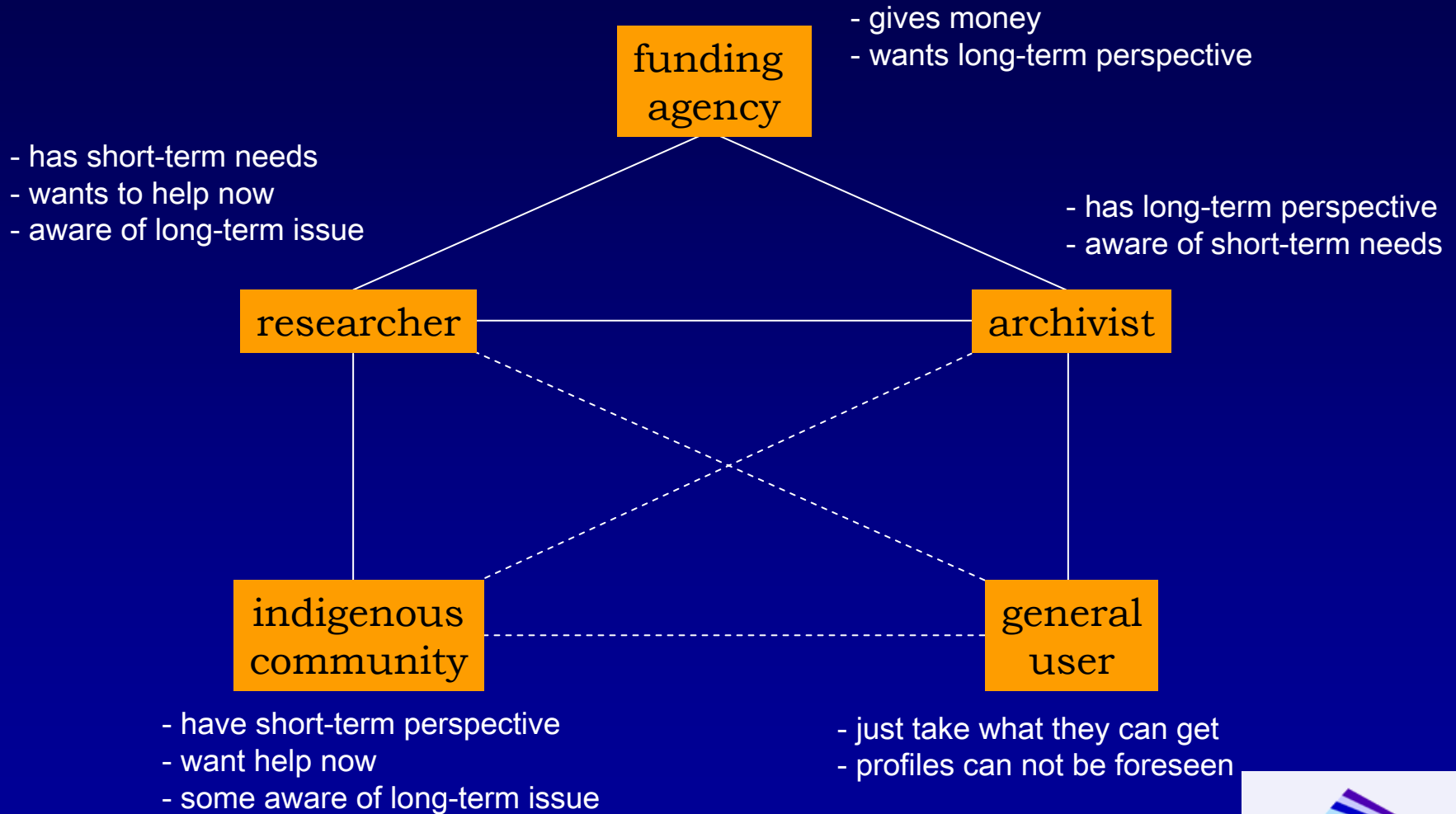


# Principal Function

- two major tasks to be solved
  - ethical obligation to document languages to preserve it for future generations
  - ethical obligation to make revitalization attempts and to help communities
- two tasks pose partly conflicting requirements
  - first has a long-term perspective
  - second has a short-term perspective



# Parties Involved



# Statement 1

Primary Goal of the DOBES Archivist:

- store all material in a safe way
- take care that it is accessible even after years

Data presentation for revitalization purposes is not the archivists task.

However, archivist has to give easy access and help the researcher or service provider.

Archivist could take the decision to collaborate with researchers to present selected material. It can be a costly enterprise.



# Statement 2

To give easy and state-of-the-art access all data has to be offered online.

To help the researchers all data has to be in well-documented formats with a chance for long-term survival.



# Pillars of Survival 1

Survival is a matter of societal acceptance.  
Relevant aspects are:

- political attitude
- attractiveness of the content after years
- involvement of recognized institutions
- cost efficiency of operation
- quality of data management and used technology

Have a responsibility to optimize in this direction -  
otherwise not worth spending the money now.



# Pillars of Survival 2

- political attitude
- attractiveness of the content after years

Hardly any influence by us.

Be wise and spread all material around the world!!



# Pillars of Survival 3

- involvement of recognized institutions

Digital Libraries are new! Don't need big stone buildings - but smart people.

Traditional libraries and museums are not ready for DL.

Need intermediate solutions.



# Statement 3

The DOBES archivist is a temporary solution until new institutions with long-term perspective have emerged.



# Pillars of Survival 4

- cost efficiency of operation

Have a fundamental problem with technology:

- today's technology is transient (6 years lifetime?)
- today's technology is not reliable

Chosen solutions:

- keep several copies at several places
- migrate all data to new storage media in time

Not really cost efficient, but automate as much as possible.



# Statement 4

## The DOBES archivist

- has three copies himself (one at another physical location)
- collaborates with a mirror site
- serves researchers with copies of the DMF
- will migrate to new storage media regularly
- will store the data this way at least for 10 years  
(no guarantee for a certain service level for users after that period)
- will help to find a solution at the end

This can be seen as a sufficient solution for the next 10 years



# Pillars of survival 5

- Quality of data management and used technology
  - degree of accessibility and interpretability
  - state and dynamics of maintenance

Accessibility and interpretability: data formats

- media data types: video, audio, photos, etc
- textual data types: annotations, lexicons, notes, etc
- various formats, especially for old data
  - archivist will convert all data to a minimal set of standard formats



# Accessibility and interpretability 2

Documentation of formats:

- archivist has to refer to documentation about standards
- archivist has to take measures when standards are replaced by new ones  
(converting is expensive, so at least archive all documentation)
- given a survival of the documentation, everyone can write appropriate algorithms
- The archiver's own formats have to be well documented (e.g. with schemas or DTDs)



# Accessibility and interpretability 3

User access:

- all data is online!
- given sufficient permissions, the user
  - can either operate directly on the files
  - or use shells offered by the archivist
- direct operation guaranteed by relying on open standards such as XML, MPEG, etc
- MD shell by IMDI tools (browsing and searching)
- annotated media by EUDICO tool set



# Data management

- resource management is important
- DOBES corpus currently 250 GB of data
- MPI corpus about 10.000 sessions
- archivist uses corpus management on conceptual level, i.e. IMDI type of metadata is used
- clear organizational principles to prevent chaos
- MD are open, access to resources dependent on communities and researcher
- no original tapes stored by archivist
- careful storing and protecting does not prevent misuse!!



# Data management: dynamics

- the data is dynamic
  - requires good version management
  - data provenance: extractions of the archive may become outdated
  - data synchronization: merge new versions of resources with the archive
- interoperability between archives
  - relevant for metadata to have a common search space
  - has a price to be paid

