Welcome to the DOBES Training Course



Paul Trilsbeek

Program

DOBES Training course October 2012 program.

	Monday 8-10	Tuesday 9-10	Wednesday 10-10	Thursday 11-10
Time				
9.00 – 10.30	Welcome Introduction to DOBES archiving	Introduction to IMDI metadata & Tools (ARBIL)	ELAN annotation tool introduction & practice	Power Management, A/V file formats, A/V compression,
10.30-11.00	Coffee break			
11.00-12.30	Audio recording, Digital audio handling, Tools and practice	IMDI metadata practice	ELAN practice	Shoebox (Toolbox) introduction & practice
12.30-13.30	Lunch break			
13.30-15.00	Video recording, Digital video handling	LAMUS archive upload and management	LEXUS lexicon tool introduction	Shoebox (Toolbox) practice
15.00-15.30	Coffee break		Coffee break	
15.30-17.00	Video tools and practice		LEXUS practice	Workflow AMS access management

17:00 Drinks

Other

- hope on interaction
- handouts, exercises
- powerpoints, manuals, other material: download package

- open for further suggestions, will try to fit it in



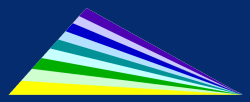
Introduction to DOBES - Overview

Sebastian Drude

with input from:
Paul Trilsbeek
Peter Wittenburg
and others

Introduction to DOBES

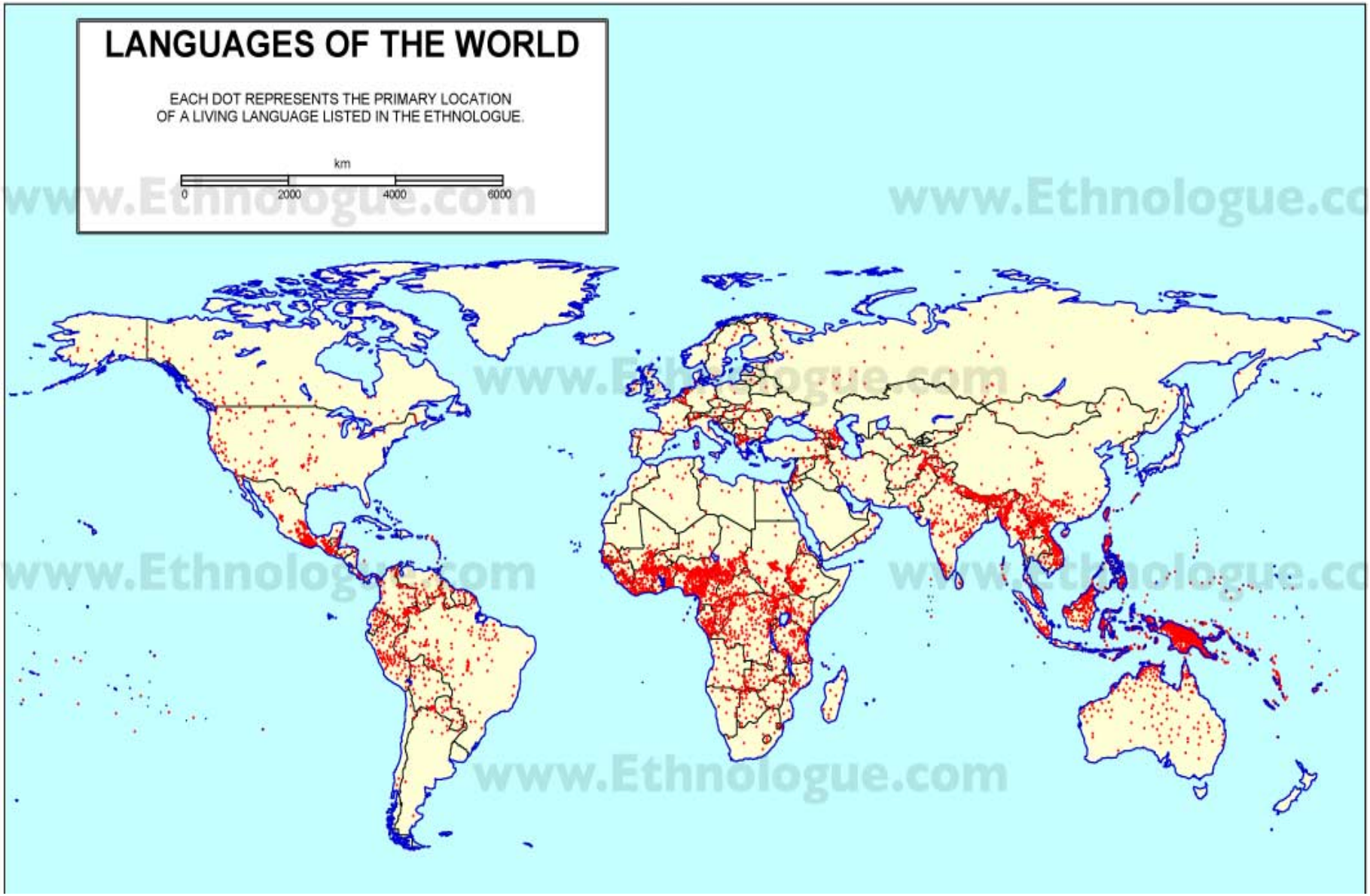
1. Background
2. Language Documentation & DOBES
3. Workflow in a DOBES project
4. Primary data and fieldwork
5. Organizational & technical tips
6. Secondary data: annotation
7. The Language Archive
8. Legal & ethical questions, agreements

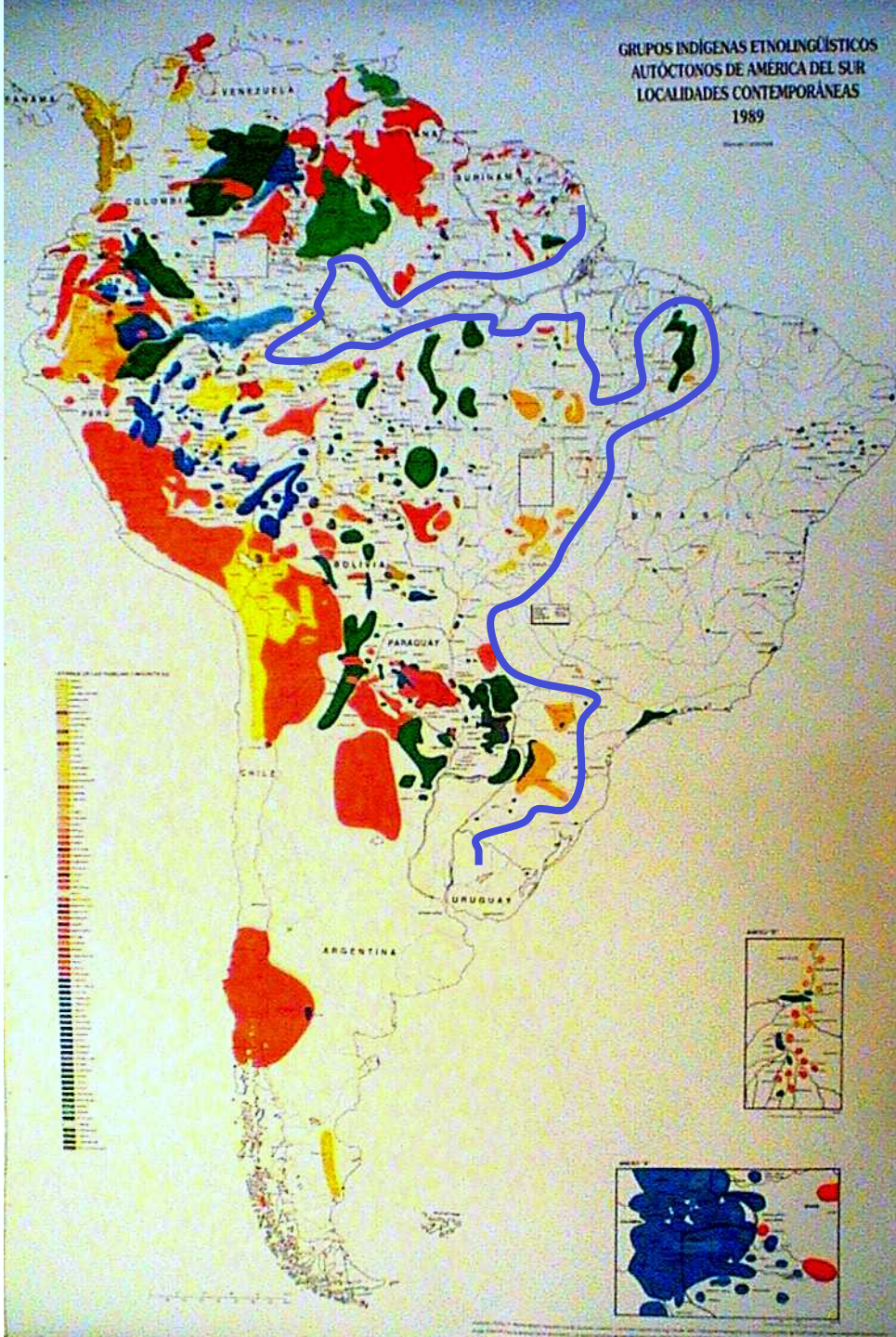
Introduction to DOBES

1. **Background**
2. Language Documentation & DOBES
3. Workflow in a DOBES project
4. Primary data and fieldwork
5. Organizational & technical tips
6. Secondary data: annotation
7. The Language Archive
8. Legal & ethical questions, agreements

1. Background: Language diversity



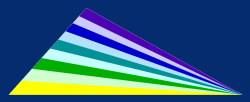

Columbia
Venezuela
Trinidad
Guayana
Surinam
French Guayana
Panama
Ecuador
Peru
Quechua
Aymara
Bolivia
Brazil
Ge languages
Tupi
Guarani
Chile
Argentina
Uruguay
South American Languages
Historical Distribution
(C) Jost Gippert, 1993-2003


GRUPOS INDÍGENAS ETNOLINGÜÍSTICOS
AUTÓCTONOS DE AMÉRICA DEL SUR
LOCALIDADES CONTEMPORÁNEAS
1989

1. Background: Language endangerment



Aikana
Ajuru
Akawaio (Brazil)
Akuntsu
Akwáwa
Amanayé
Anambé
Apalai
Apiaká
Apinajé
Apurinã
Arapáso
Arara do Pará
Arára
Shawãdáwa
Araweté
Arikapu
Aruá
Ashaninka
(Brazil)
Asurini do Xingu
Aurê-Aurá
Ava-Canoeiro
Aweti
Bakairi
Banawá Yafi
Baníwa do Içana
Bara (Brazil)
Barasana (Brazil)
Baré (Brazil)
Bororo

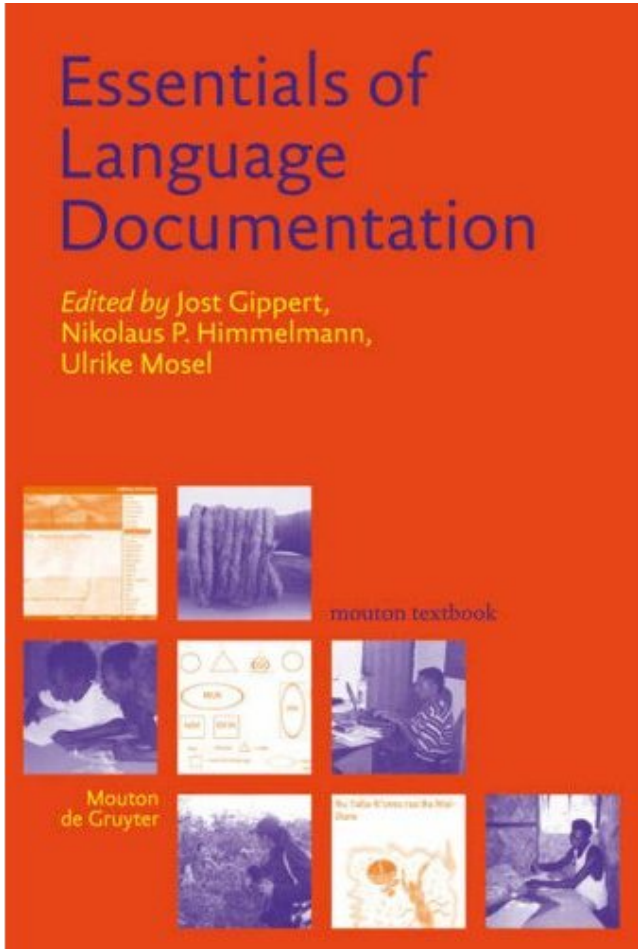
Introduction to DOBES

1. Background
2. **Language Documentation & DOBES**
3. Workflow in a DOBES project
4. Primary data and fieldwork
5. Organizational & technical tips
6. Secondary data: annotation
7. The Language Archive
8. Legal & ethical questions, agreements

2. Language Documentation

- New academic area (Himmelmann 1998)
- Triggered by language endangerment, enabled by technical change
- In addition to the "Boas'ian triad" (grammar, dictionary, text collection): **corpora of annotated multimedia-data**
- Focus on **natural language use**
- **Multi-purpose**:
  - language typology, linguistic analysis, ...
  - anthropology, ethno-musicology, ethno-biology...
  - oral history, archeology, lang. revitalization,...
- Some future uses may be unknown

2. Language Documentation: Sources



LANGUAGE DOCUMENTATION & CONSERVATION



rnld
Resource Network for Linguistic Diversity

EMELD

The E-MELD School of
Best Practices in Digital Language Documentation

2. Language Documentation: Sources

- Gippert, J.; N.P. Himmelmann; U. Mosel, Eds. (2006). *Essentials of Language Documentation*. Berlin, New York, Mouton de Gruyter.
- ***Language documentation and conservation***. Honolulu, Hawaii, U. o. Hawaii P. (1.2007 -). http://www.nflrc.hawaii.edu/ldc/default.html
- ***Language documentation and description***. London, SOAS, Univ. of London. (1.2003 -).
- http://www.rnld.org
- http://emeld.org/school of best practice
- http://www.mpi.nl/dobes
- http://tla.mpi.nl/

2. Language Documentation: Projects















AIATSIS
AUSTRALIAN INSTITUTE OF ABORIGINAL AND TORRES STRAIT ISLANDER STUDIES





THE HANS RAUSING
Endangered Languages Project







Documenting Endangered Languages (DEL)
data, infrastructure and computational methods



2. Language Documentation: Projects

Funding programs, centres, projects & initiatives

- **DOBES & TLA (D&NL)**
- E-MELD (US)
- AILLA (US)
- PARADISEC (AU)
- **HRELP (UK)**
- ANLC & **DEL (US)**
- LACITO (FR)
- AIATSIS (AU)
- AMPM (NZ)
- etc.

- DELAMAN: “Living Archive” Initiative
- OLAC: Open Language Archives Community
- CLARIN research infrastructure project
- EUDAT
- DASISH
- …

2. DOBES: General

- **DO**kumentation **BE**drohter **S**prachen
  Documentation of Endangered Languages
- Funded by Volkswagen**Stiftung**
- Two major pillars:

1. ~65 language documentation projects
   - Focus on data (linguistic analysis, revitalization and other activities welcome but additional)
   - Language documentation in cultural context
   - Partnership with speech community
2. Creating an organized, accessible and **persistent** archive

2. DOBES: Field sites



2. DOBES: History

- Interest in “endangered languages” since 1991/92, LSA symposium etc.
- DGfS & VWS summer-school in Köln 1993
- Conceptional preparation until 1999
- Pilot Phase 2000/2001:
  5+3 individual projects &
  ‘TIDEL’ (TG @ MPI-PL, today: “TLA”)
- 2001–2012: calls for main phase,
  each year around 5–8 new projects
- End of programme ~2015

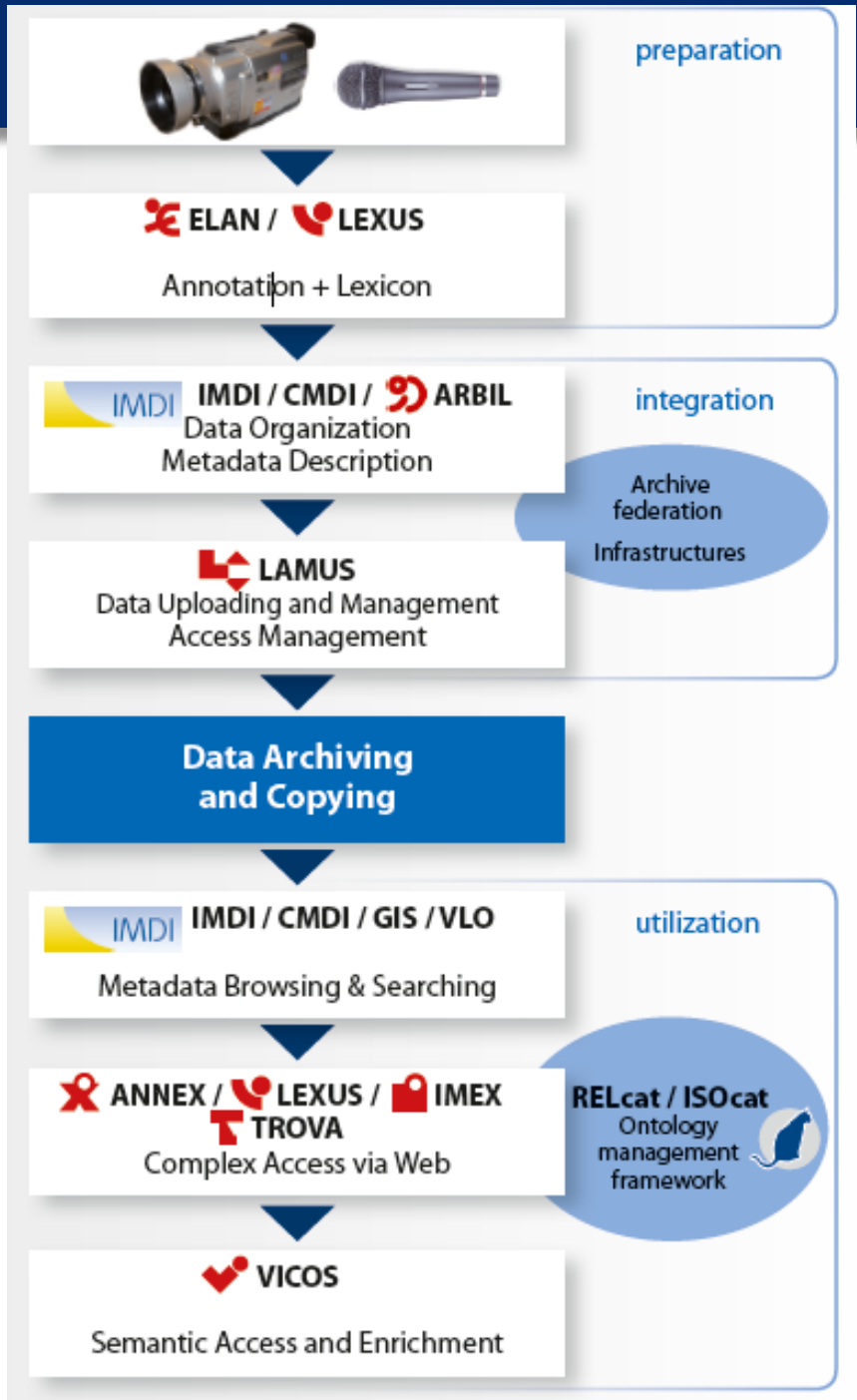
1. Background
2. Language Documentation & DOBES
3. **Workflow in a DOBES project**
4. Primary data and fieldwork
5. Organizational & technical tips
6. Secondary data: annotation
7. The Language Archive
8. Legal & ethical questions, agreements

3. Workflow: overview

1. Recording of speech events and culture
2. “Digitalization” of recordings
3. File naming and local organization
4. Creation of metadata
5. Annotation 1: Transcription
6. Annotation 2: Translation
7. Annotation 3: Glossings & Lexicon
8. Annotation 4: Advanced annotation
9. Corpus organization & archiving
10. Administration: Access rights etc.

3. Workflow: Tools

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



Program

Program

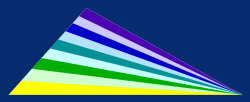
DOBES Training course October 2012 program.

	Monday 8-10	Tuesday 9-10	Wednesday 10-10	Thursday 11-10
Time				
9.00 – 10.30	Welcome Introduction to DOBES archiving	Introduction to IMDI metadata & Tools (ARBIL)	ELAN annotation tool introduction & practice	Power Management, A/V file formats, A/V compression,
10.30-11.00	Coffee break			
11.00-12.30	Audio recording, Digital audio handling, Tools and practice	IMDI metadata practice	ELAN practice	Shoebox (Toolbox) introduction & practice
12.30-13.30	Lunch break			
13.30-15.00	Video recording, Digital video handling	LAMUS archive upload and management	LEXUS lexicon tool introduction	Shoebox (Toolbox) practice
15.00-15.30	Coffee break		Coffee break	
15.30-17.00	Video tools and practice		LEXUS practice	Workflow AMS access management

17:00 Drinks

3. Workflow: overview

1. **Recording of speech events and culture**
2. "Digitalization" of recordings
3. **File naming and local organization**
4. Creation of metadata
5. Annotation 1: Transcription
6. Annotation 2: Translation
7. Annotation 3: Glossings
8. Annotation 4: Advanced annotation
9. Inclusion into corpus & archive
10. **Administration: Access rights etc.**

Introduction to DOBES

1. Background
2. Language Documentation & DOBES
3. Workflow in a DOBES project
4. **Primary data and fieldwork**
5. Organizational & technical tips
6. Secondary data: annotation
7. The Language Archive
8. Legal & ethical questions, agreements

4. Primary data and **fieldwork**

- Fieldwork is a fundamental aspect of LD
- Good fieldwork is an art (methodology, tradition, vocation)
- Learn from fieldworkers – highlights, and especially failures
- Respect, honesty, humbleness, humor
- Speaker not as “objects”, but partners
- Planning of documentation
- Payment
- Informed consent

4. Primary data and fieldwork: Ethnography

Ethnography is an important tool for language documentation, because:

- E. identifies important oral texttypes and genres (ethnography of speech)
- E. gives a rich context for speech acts
- E. helps identifying key concepts

⇒ The more of the culture is understood, the better the documentation

⇒ The best language documenters are the speakers themselves ("the community"), also against the **observer's paradox**

4. Primary data and fieldwork: who









4. Primary data and fieldwork: who





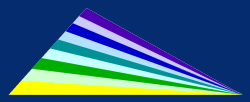
4. Primary data and fieldwork: who

- Research 'with' vs research 'about'
- Who is 'the community'?
- Who participates? Which roles? (Speaker, semi-spaker, leaders, teachers and pupils, etc.)
- Advantages: more productive, representative, inside-view, effective
- Challenges / limits:
  - be flexible, accept new goals – deliverables
  - training – time factor
  - internal politics

4. Primary data and fieldwork: what

- NATURAL DISCOURSE, DIALOGUES, REAL EVENTS
- Basic word lists, kinship terms
- The body: Parts, processes & life cycle (conception, soul, menstruation, reclusion m/f, pregnancy, birth, marriage, grandchildren & growing old, death, etc.)
- "Religion": Natural and supernatural beings, diseases, curing, blessings/curses, witchcraft, etc.
- Material culture / production of objects: Processes, how-tos, names and symbolic meanings, ...
- Subsistence: Processes, taxonomies, cultivation and use of plants, calender, astronomy...
- Mythical tales: about origin of world, entities, practices, objects, peoples, dead, language...

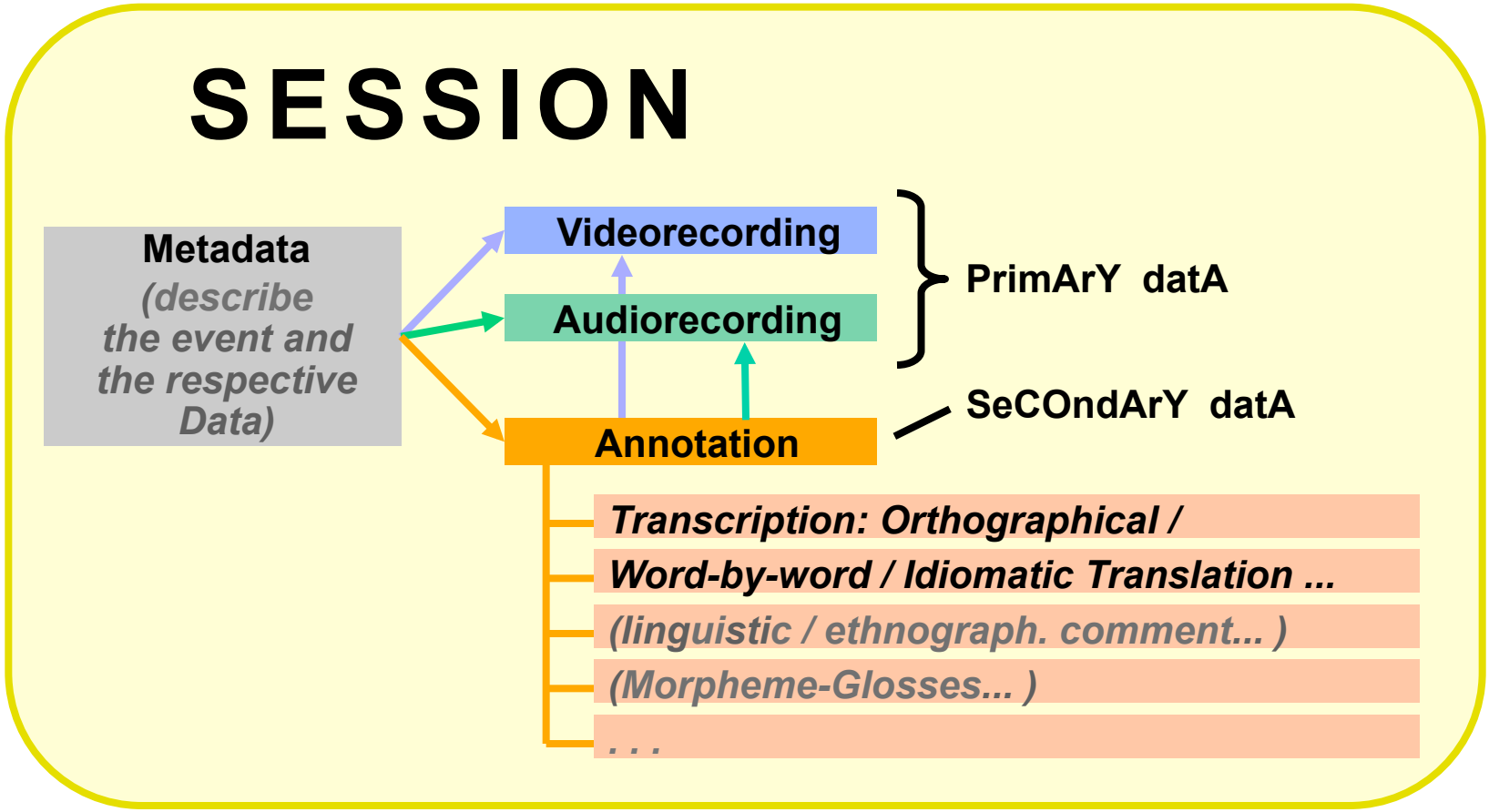
→ **LOOK AT THE EXAMPLES IN THE ARCHIVE**

Introduction to DOBES

1. Background
2. Language Documentation & DOBES
3. Workflow in a DOBES project
4. Primary data and fieldwork
5. **Organizational & technical tips**
6. Secondary data: annotation
7. The Language Archive
8. Legal & ethical questions, agreements

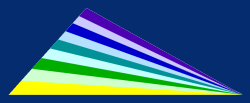
5. Organizational & technical tips



5. Organizational & technical tips

Typical situation:

- Large and growing number of files
- Many have to be checked or processed
- Number of sessions grows, potentially each with different workflow status
- Many different types of related files
- Different versions of the same file
- Different computers – individual laptops, in the field, office, etc.

5. Organizational & technical tips

Question:

How not to get lost in so many files, and different versions of files, on different computers and DVDs, HDs?

People tend to underestimate the importance of **consistent solutions**

THINK FIRST!

PLAN YOUR ORGANIZATION AND WORKFLOW!

WRITE IT DOWN!

5. Organizational & technical tips

Recommendations:

- Consistent naming of files
- ONE identical directory structure (‘tree’) for the project on all computers
- Create a virtual drive for this ‘tree’
- Use an adequate file manager (Windows: Total Comander)
- Regular backups and synchronizations

5. Organizational & technical tips

- All related files (e.g., same session) should have similar file names, starting identically
- All files that contain data of the same event (primary and derived data in different format) should have the same name, differing only in the ***extension*** (**.wav** – **.mpg** – **.sft** – **.eaf** …)
- Automatical sorting of files should render a useful grouping and order of files, e.g.:
- Date & time:
  year-month-day hour:minute:sec.frame/msec
  2011-09-24 14:35:12.10 –or:–
  20110924:14351210

5. Organizational & technical tips

Recommendation (solution at Museu Goeldi):

AWE_20090526_SD_157_BasicList400_birds.wav

- **AWE** : ISO-639 language code (if appropriate)
- **20090526** : Date of event (recording), no spaces
- **SD** : Code of researcher (optional)
- **157** : Code of speaker (here anonymized, can be several)
- **BasicList400** : Label for (content of) SESSION
- **birds** : Label for (content of) FILE
- **wav** : extension (type / format of file)

5. Organizational & technical tips

Example for a narrative:

AWE-200208-157-kujakyt-3-end.wav

- **AWE** : ISO-639 language code for “Awetí”
- **2002**[**08**] : ***Day (and even month) is optional***
- [**SD**] (“missing”) : ***Code of researcher is optional***
- [**157**] : Code of speaker (***can be several, or “men” etc.***)
- [**3-end**] : ***If order is relevant, start with a number***

Minimal example:

AWE-1972-elicit-3.wav

Naming of *raw data* files:

AWE-20020807-SD-0471of1638.wav

- **AWE** : ISO-639 language code (if appropriate)
- **20020807** : Date of event (recording)
- [**SD**] : Code of researcher or device (if necessary)
- **0471** : ***Original* number** of file / track
- **of** : Short word indicating that what follows is …
- **1638** : **Total number** of tracks (on device & field trip)
- **wav** : extension (type / format of file)

5. Organizational & technical tips

Challenge of ***sustainability***:

- ***Physical level***: limited lifetime of carriers
  → constant copying & replacement of carriers
- ***Logical level***: limited lifetime of formats





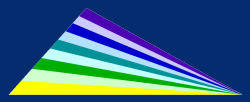
0 years 200 years 500 years 1000 years 2000 years

electronic media paper clay tablets

  → adherence to standards:
  Unicode, XML, open formats
  → constant updating of encodings
  → Ca[illegible]h transformations
  lo[illegible]dings
  artefacts can be introduced
  → Always provide **provenance information**





Introduction to DOBES

1. Background
2. Language Documentation & DOBES
3. Workflow in a DOBES project
4. Primary data and fieldwork
5. Organizational & technical tips
6. **Secondary data: annotation**
7. The Language Archive
8. Legal & ethical questions, agreements

6. Secondary data: annotation

Data is always data FOR something, or at least OF something – usually it is ***a systematic representation of*** physical objects, states and events

In linguistics, *primary data* is a direct rendering or result of *communicative (speech) events*, for instance a ***written text*** or, in particular, an ***audio/video recording*** of a speech event

6. Secondary data: annotation

Annotation of data is ***a symbolic representation*** of properties of the object/state/event represented in the (primary) data

⇒ Annotation is also (secondary) data

In linguistics, the most common and basic types of annotation are a ***transcription*** and a ***translation*** of the linguistic expressions represented in primary data (e.g., in an audio / video recording)

This is the agreed minimum in DOBES

6. Secondary data: annotation

Global or *holistic annotation* represents properties of the object/state/event as a whole and is part of the *metadata*

Unit-oriented annotation refers to specific parts of the data, in linguistics, in particular, to utterances of individual sentences or words or sounds etc.

We speak of *individual* ***annotations*** (plural)

6. Secondary data: annotation

If unit-oriented annotation is directly based on primary data then it is *secondary data*

Annotation based on secondary data would be tertiary data, and so forth recursively

In sum, all unit-o. annotation is *derived data*

There are other types of derived data that are systematizations and abstractions (lexicon...)

Ethnographic & linguistic description & analysis and other academic work are not “data” per se, but are *based on* data → references!



6. Secondary data: annotation

Annotation of a media file is *time-aligned anotation* if each piece of annotation is explicitly associated with the corresponding *chunk* (***time-span, segment***) of the media file

This is usually done by using the time position of the ***start and end points*** of the respective chunk, the *time marks*

6. Secondary data: annotation

Annotations differ according to the *types of properties* of the object/state/event that are represented in the annotation

In linguistics, annotations can be *phonetic, phonological, morphological, syntactic, semantic, pragmatic*, (possibly others), and on each level they can focus on the ***units***, or on ***structures*** of units, or on ***relations*** that hold among units, etc.

On each level possibly also: comments, author attribution, date / log

6. Secondary data: annotation

In Linguistics:

Basic annotation: only *transcriptions*, *translations* and perhaps *notes*, on a sentence level **(DOBES minium)**

Basic glossing: additionally information on *individual morphs*: a *gloss* (indication of meaning or function) and perhaps a part-of-speech tag **(recommended)**

Advanced glossing: one or several of *additional levels*, from phonetic to pragmatic (for instance, a *prosodic transcription*, or annotation of the *syntactic structure*, of *grammatical relations*, etc.) **(recommended for a smaller sample)**

A syntactic glossing table

I	1	2	3
II	[di.	ʔʊn̩y.bɐ.zɪç.tʰɪç̊n̩	pʁo.blee.mə]
III	L	H_f M_f L H H H	H H_r L_f
IV	/ ˌdii. /	/ ˈʔun.ˌʔyy.bə.r-ˌziX.t-ˌliχ̟ə.n /	/ pro.ˈblee.mə /
V	L	H L L H H H	H **H_r** L
VI	*die*	*unübersichtlichen*	*probleme*
VII	DefArt	Adj	Sub Neut
VIII	Nom Pl Unm_G Str	Unm_C Pl Unm_G Wk	Nom Pl Unm_D
IX	Nom Pl Def	‘involved’	‘problem’
X	1,3: Nf 2: Nf **1**,2,**3**: NGr		
XI	mod: 2 1,3		
XII	Die unübersichtlichen Probleme.		
XIII	E: The involved problems.		

A morphological glossing table

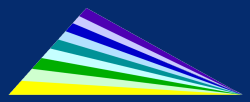
I	1	2	3	4	5
II	/ ʔun.	ʔyy.bə.r-	ziX.t-	liX̣	ə.n /
III	H	H,L L	H,L	H,L	L
IV	/ ˈʔun. /	/ ˈʔyy.bə.r /	/ ˈziX.t /	/ ˌliX. /	/ ə.n /
V	H	H,L L	H,L	H,L	L
VI	*un*	*über*	*sicht*	*lich*	*en*
VII	$Pref_i$	$PrepSt_j$	$SubSt_k$	$SubSt_l/AdjSt_m$	$AdjFlex_n$
VIII	–	–	–	–	–
IX	not	‘over’	‘view’	suitable-for	Unm_C Pl Unm_G Wk
X	1: Af 2: Stf	3: Stf 4: Af	5: Af 2,**3**: Stf	**2,3**,4: Stf 1,**2,3,4**: Stf	**1,2,3,4**,5: StGr
XI	m-mod: 2 3	m-qual: 4 2,3 3	m-mod: 1 2,3,4	m-qual: 5 1,2,3,4	1,2,3,4
XII	unübersichtlichen				
XIII	‘involved’				

6. Secondary data: annotation tools

	Transcriber	ELAN	Toolbox	FLEX
Complexity	Easy	Complex, w. easier modes	Complex to configure	Complex
Audio	Yes	Yes	No (can play)	No
Video	No	Yes	No	No
Tiers	1 per speaker	Unlimited	Unlimited	Fixed: 8
Lexicon interop., automatic glossing	No	No (is planned)	Yes	Yes
Unicode	No input	Yes	Yes	Yes
Data format	Simple XML	Compl. XML	Faulty TXT	XML database
Interoperability	Good	Fair	Good	Bad
User community / support	Small?, no support?	Large, good support	Large, fair support	Small, good support
Life cycle	Old (but new version 2011)	Constantly developed	Not officially supported, old	New, being developed

6. Secondary data: annotation tools

	Transcriber	ELAN	Toolbox	FLEX
Complexity	Easy	Complex with easier modes	Complex to configure	Complex
Audio	Yes	Yes	No (can play)	No
Video	No	Yes	No	No
Tiers	1 per speaker	Unlimited	Unlimited	Fixed: 8
Lexicon interop., automatic glossing	No	No (is planned)	Yes	Yes
Unicode	No input	Yes	Yes	Yes
Data format	Simple XML	Compl. XML	Faulty TXT	XML database
Interoperability	Good	Fair	Good	Bad
User community / support	Small?, no support?	Large, good support	Large, fair support	Small, good support
Life cycle	Old (but new version 2011)	Constantly developed	Not officially supported, old	New, being developed

6. Secondary data: sustainable annotation

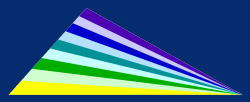
Standards and recommendations:

- Open (and free) formats
- Non-binary (text-based)
- **XML** (consistent, with suitable schema)
- Encoding: **UNICODE**
- Be explicit about conventions, specify:

– Which tier contains what information?

– Standards for labels for tiers!

– Meaning of abbreviations and symbols?

Introduction to DOBES

1. Background
2. Language Documentation & DOBES
3. Workflow in a DOBES project
4. Primary data and fieldwork
5. Organizational & technical tips
6. Secondary data: annotation
7. **The Language Archive**
8. Legal & ethical questions, agreements

7. The Language Archive

- The DOBES archive is only one part of the large archive of language data at the MPI
- Establishing **“The Language Archive”** aims at a long-term perspective for our activities and hence, at sustainable archive

State of Archive:

- about 80 TB of well-de d resources
- about 20.000 hours o d audio/video recordi
- about 73.000 se
- about 4.5 milli ments
- about 47 lexica



7. The Language Archive

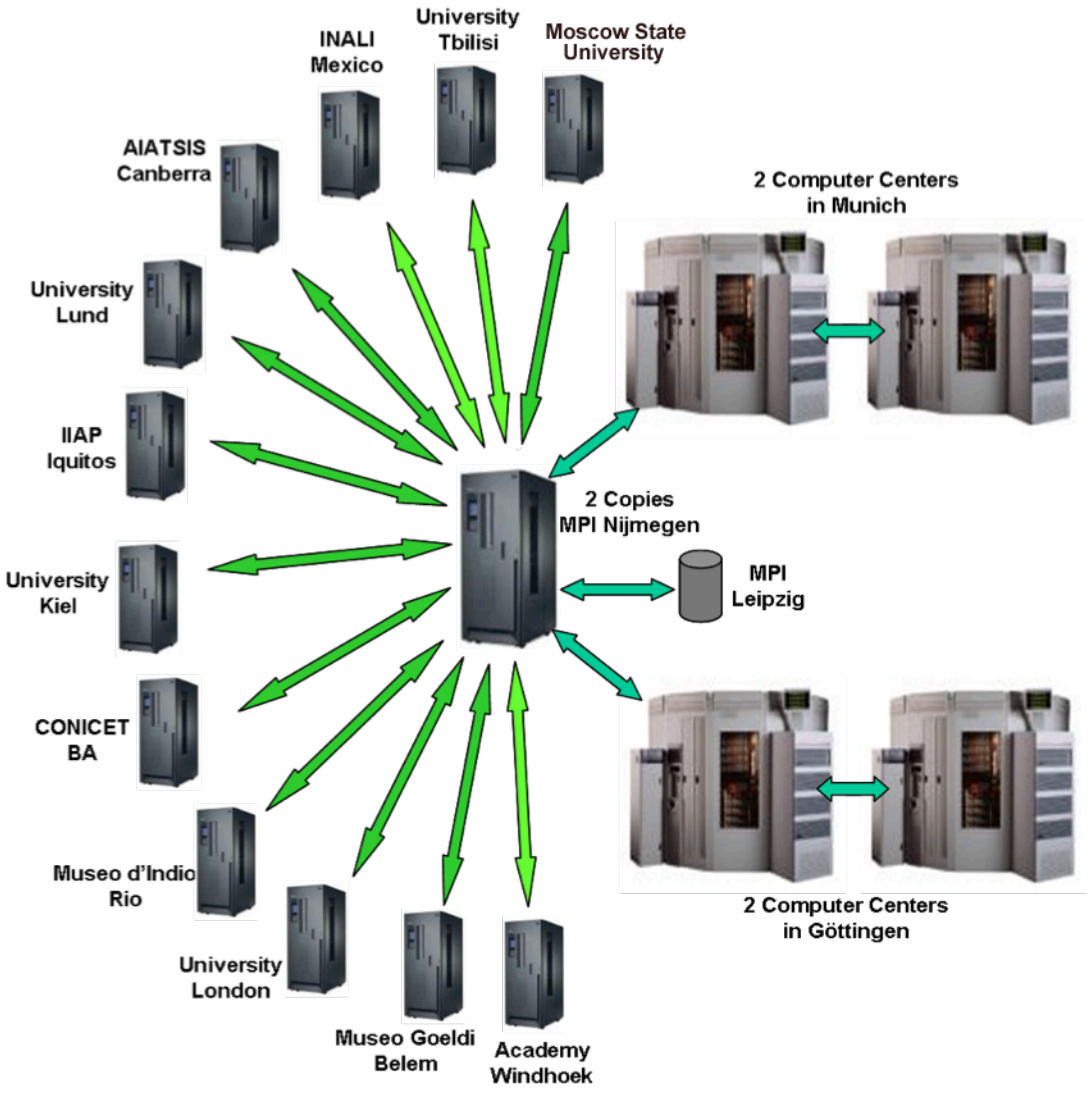
- Unit within the Max Planck Institute for Psycholinguistics Nijmegen, split off from MPI Technical Group on 1st of September 2009
- Now a combined “project” funded by the MPI for Psycholinguistics, the Max Planck Society, the Berlin-Brandenburg Academy of Science and the Royal Netherlands Academy of Arts and Sciences
- Additional funding from participation in numerous projects (EC, VolkswagenStiftung, DFG, BMBF,...)
- Officially launched on October 11 2011 in Berlin
- Headed by Wolfgang Klein (Scientific Director) and Peter Wittenburg (Head of Unit) for the coming 1 ½ years
- Currently about 27 employees + student assistants

7. The Language Archive: Data copies

Six automatic full copies at three locations in NL and in Germany

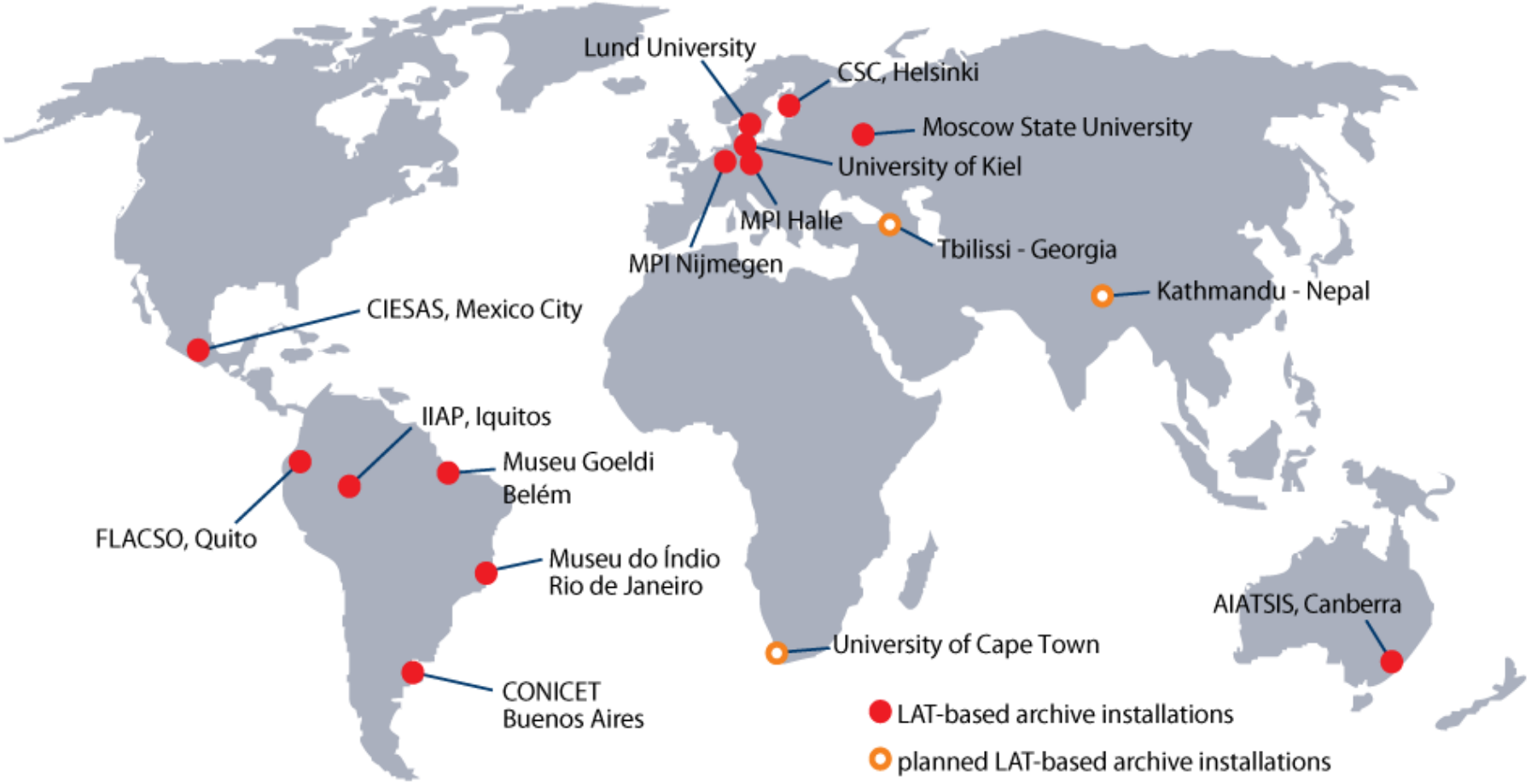
Institutional guarantee for bitstream-preservation by the MPG for 50 years

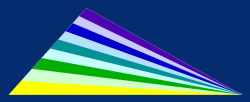
Growing number of archives use LAT



7. The Language Archive: “regional” archives

Language Archiving Technology (LAT) archives worldwide



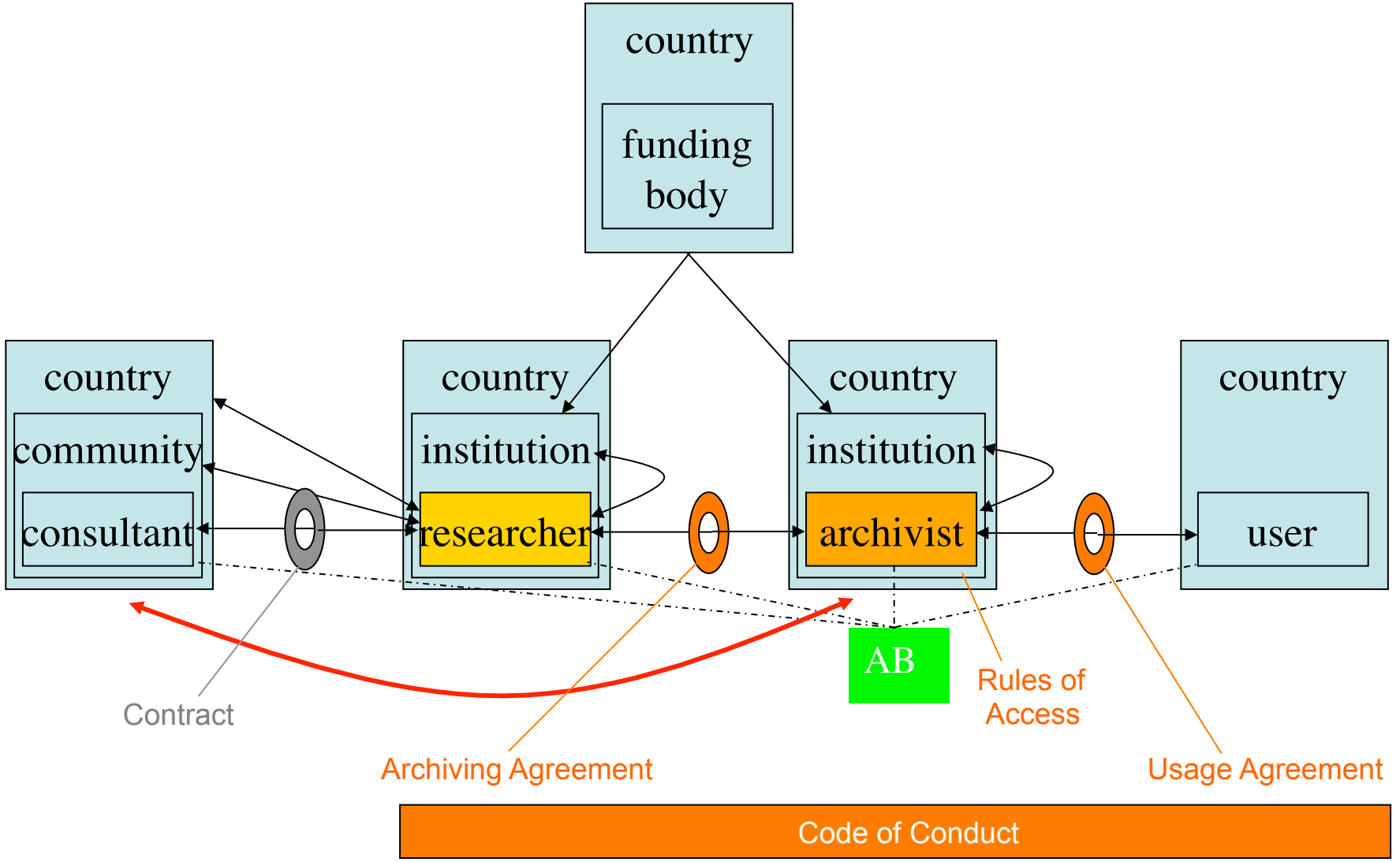
Introduction to DOBES

1. Background
2. Language Documentation & DOBES
3. Workflow in a DOBES project
4. Primary data and fieldwork
5. Organizational & technical tips
6. Secondary data: annotation
7. The Language Archive
8. **Legal & ethical questions, agreements**

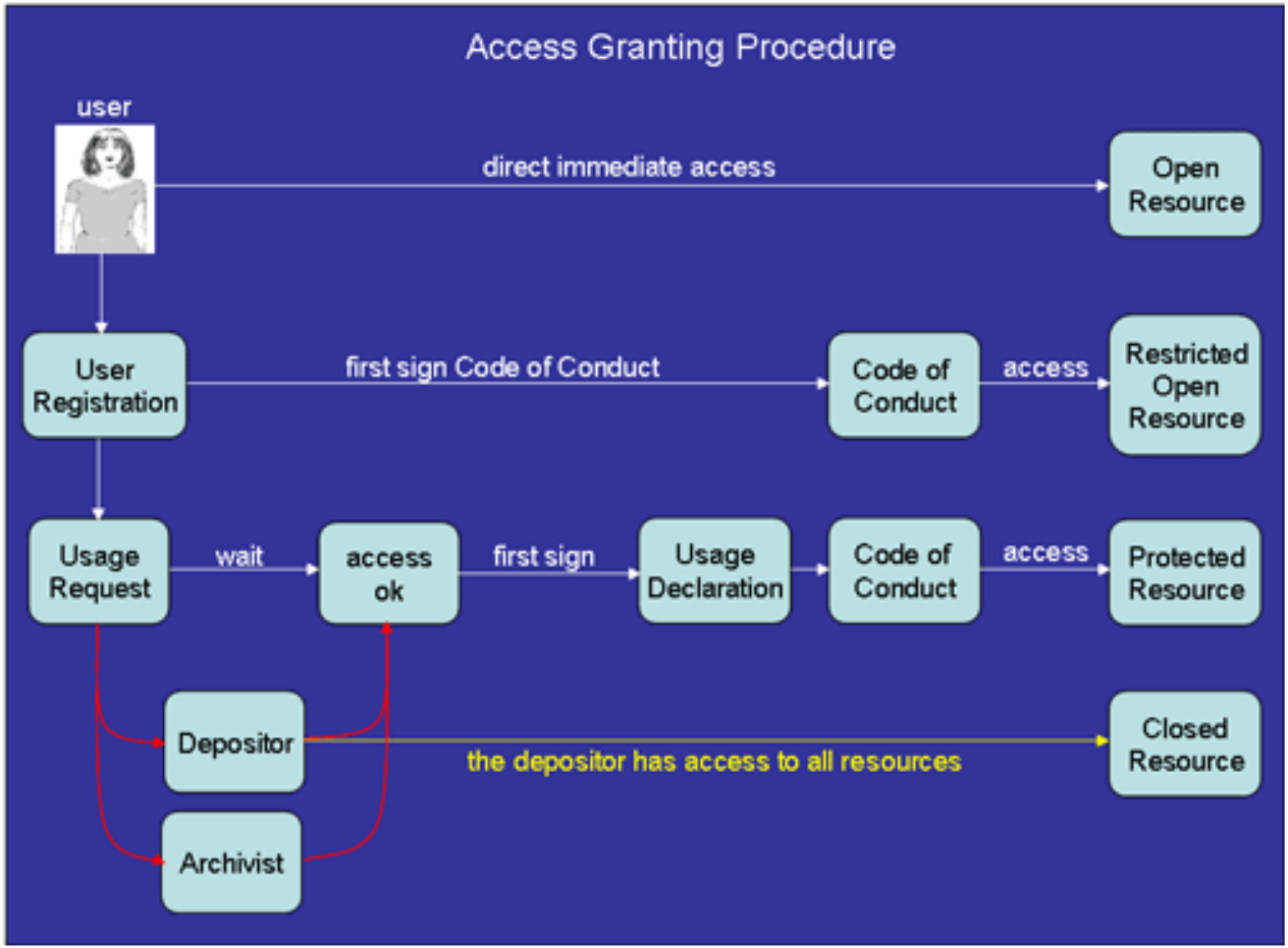
8. Legal & ethical questions, agreements

- 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
- **Who's data is "my" data??**
- TLA only claims the right to archive!

8. Legal & ethical questions, agreements



8. Legal & ethical questions, agreements



8. Legal & ethical questions, agreements

- Think ahead of anonymization
- Make clear agreements with speakers
- Plan and **describe** the access rules:

– according to topic (culturally sensitive,...)
– according to file type (audio, video, picture, annotation – metadata is open!)
– according to user groups & aims (speakers, colleagues, scientists, press, public,...)
– according to time scale (during project, three years after project, dead of speakers, unavailability of researchers)

DOBES & TLA Contact

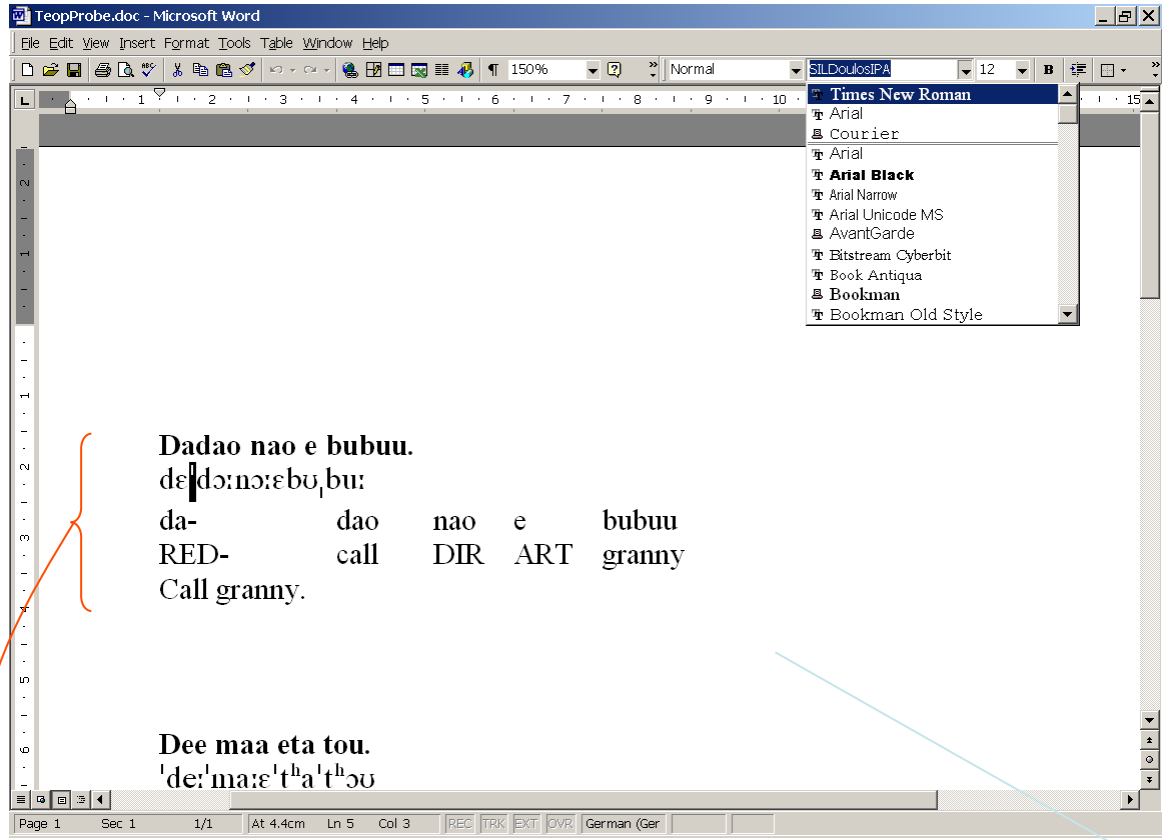
Contact persons for various aspects:

Formal DOBES Responsibility:	Peter Wittenburg
Team Interaction/Archiving:	Paul Trilsbeek Alexander König
Scientific Coordinator:	Sebastian Drude
Digitization:	Nick Wood Chris Haskett
Software Aspects:	Han Sloetjes
Electronic-A/V-Aspects:	Gerd Klaas

Important email/web addresses:

DOBES website:	www.mpi.nl/DOBES
DOBES email List:	dobeslist@mpi.nl
Data Delivery:	dobesarc@mpi.nl
TLA Web-Site, with tools:	tla.mpi.nl

Problem: Text conversion



Word document with IPA Kiel Font

“odd” characters

character mapping tables

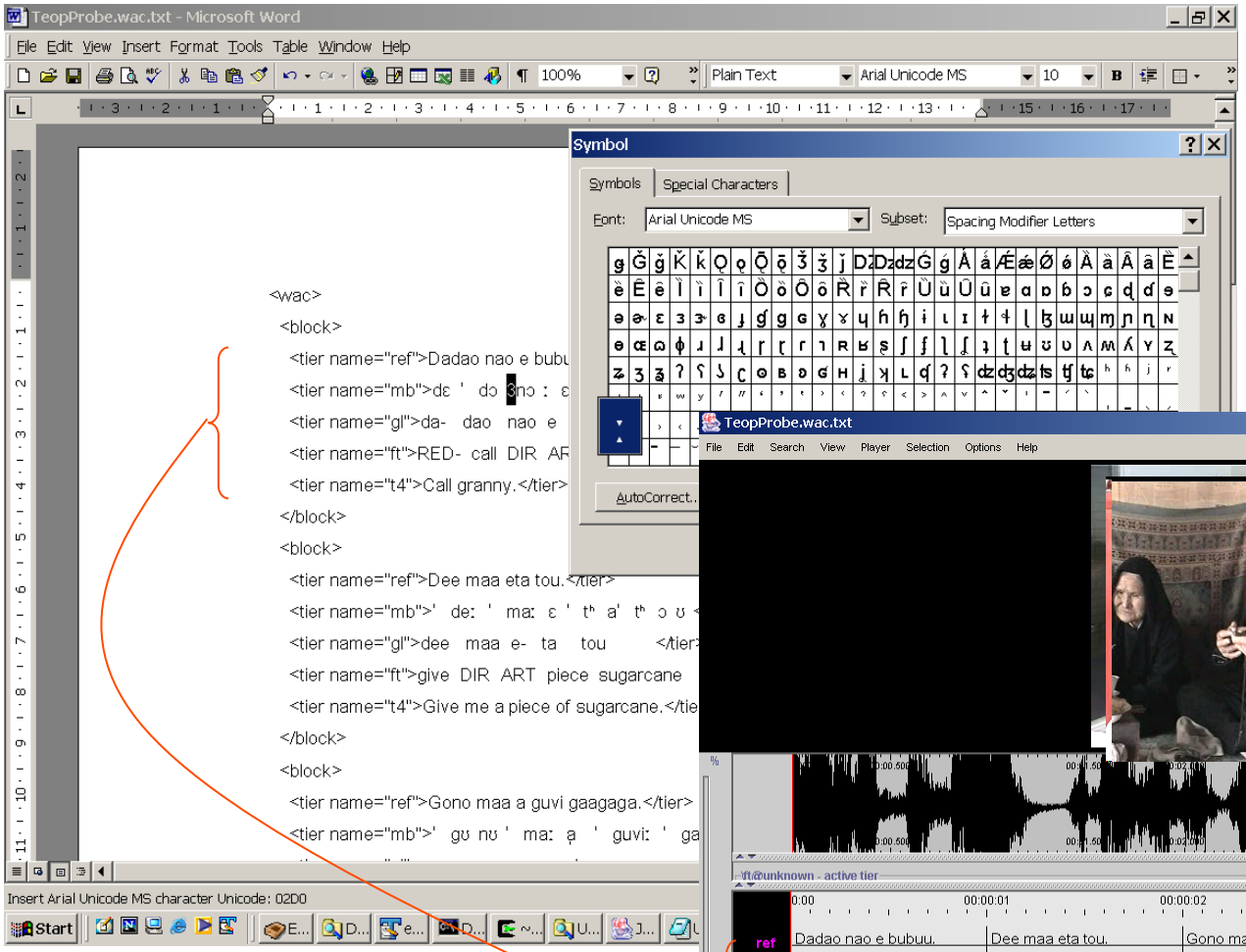
ADL

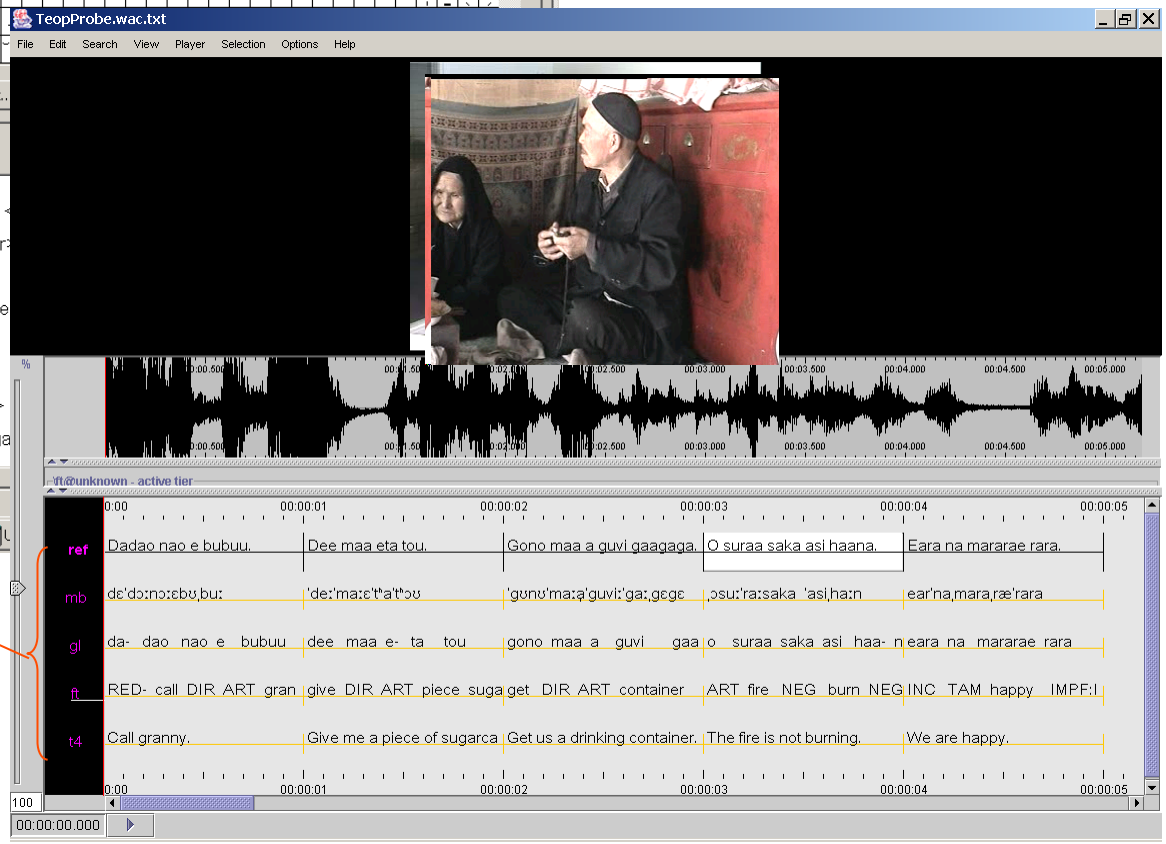
BLOCK =
TIER ortho,
TIER phonem,
TIER morphem,
TIER gloss,
TIER trans

Word document structure description

Word converter

Problem: Text conversion

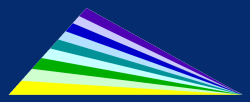




first XML file

all very expensive

then input to ELAN and time alignment etc

The End

Thanks for your attention.