Intl' DH Congress 2012
2012-07-19
Hamburg

Language Documentation and Digital Humanities: The (DoBeS) Language Archive

Sebastian Drude, Paul Trilsbeek, Daan Broeder

The Language Archive – Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

Language Documentation & DH

1) Language diversity and endangerment

2) Language Documentation

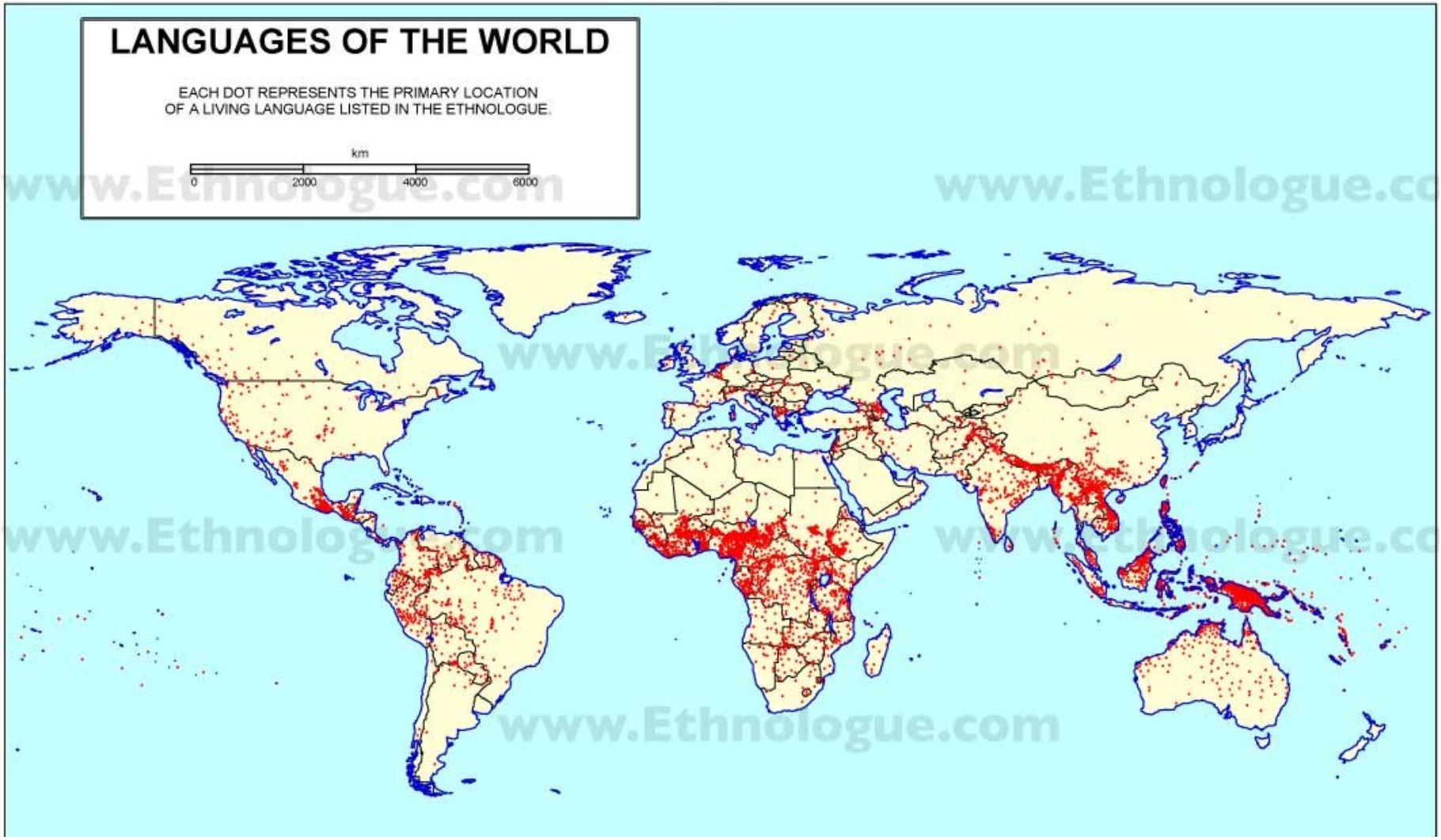
3) The DOBES Programme

4) The Language Archive

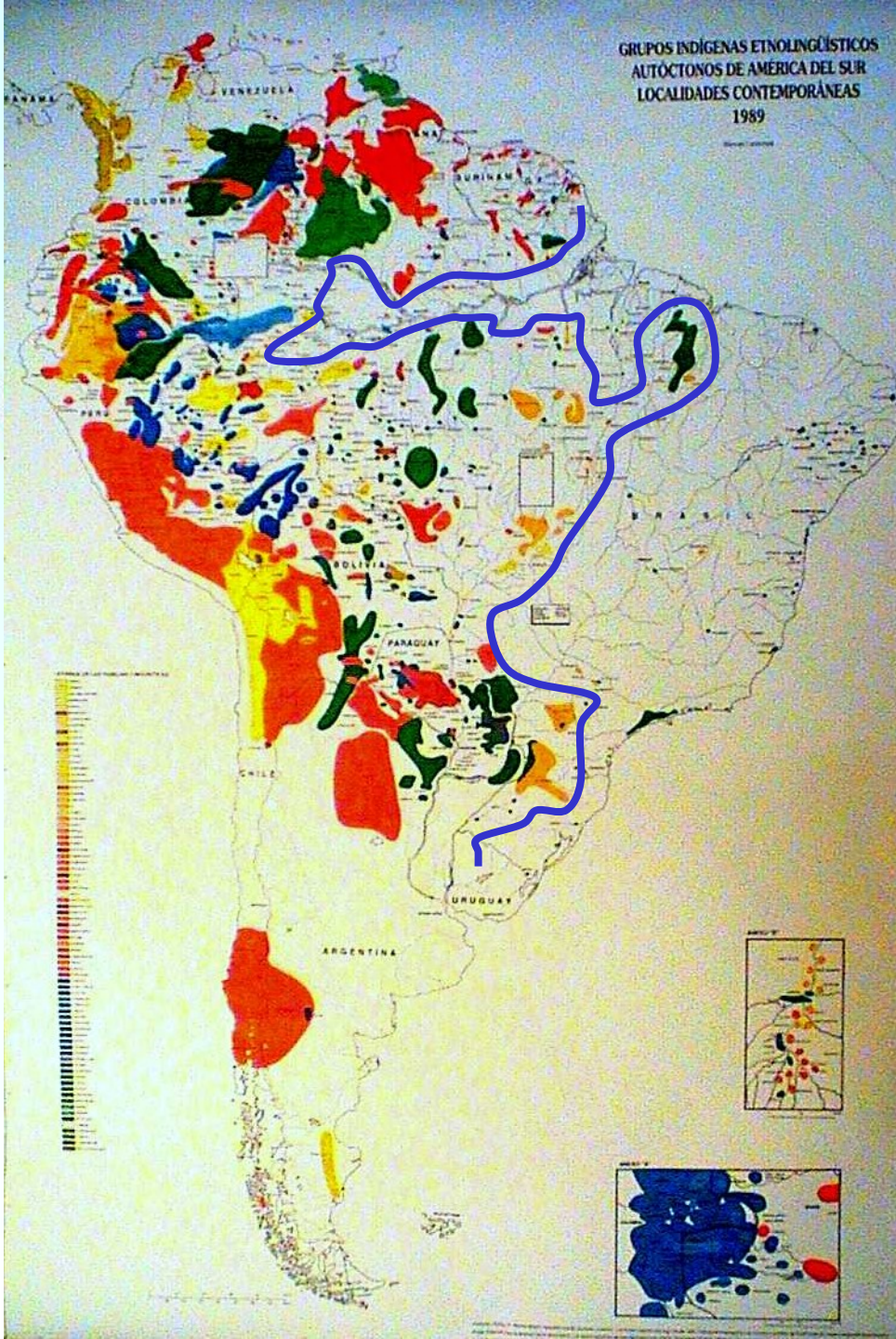
5) Relation to the Digital Humanities

1) **Language diversity and endangerment**
2) Language Documentation
3) The DOBES Programme
4) The Language Archive
5) Relation to the Digital Humanities

1) Linguistic diversity




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


GRUPOS INDÍGENAS ETNOLINGÜÍSTICOS
AUTÓCTONOS DE AMÉRICA DEL SUR
LOCALIDADES CONTEMPORÁNEAS
1989
VENEZUELA
COLOMBIA
PERU
BOLIVIA
PARAGUAY
CHILE
URUGUAY
ARGENTINA

1) Language endangerment

Degree of endangerment	Intergenerational Language Transmission
safe	language is spoken by all generations; intergenerational transmission is uninterrupted *>> not included in the Atlas*
vulnerable	most children speak the language, but it may be restricted to certain domains (e.g., home)
definitely endangered	children no longer learn the language as mother tongue in the home
severely endangered	language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves
critically endangered	the youngest speakers are grandparents and older, and they speak the language partially and infrequently
extinct	there are no speakers left *>> included in the Atlas if presumably extinct since the 1950s*

From: UNESCO Atlas of the World's Languages in danger

1) 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

1) Language diversity and endangerment

2) Language Documentation

3) The DOBES Programme

4) The Language Archive

5) Relation to the Digital Humanities

2) Language documentation

- New subfield of linguistics (Himmelmann 1998)
- Triggered by language endangerment, enabled by technical / digital revolution
- 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

DoBeS

EMELD

OPEN LANGUAGE ARCHIVES

PARADISEC

AIATSIS
AUSTRALIAN INSTITUTE OF ABORIGINAL AND TORRES STRAIT ISLANDER STUDIES

delaman

THE HANS RAUSING Endangered Languages Project

LACITO
Civilisations Langues Tradition Orale

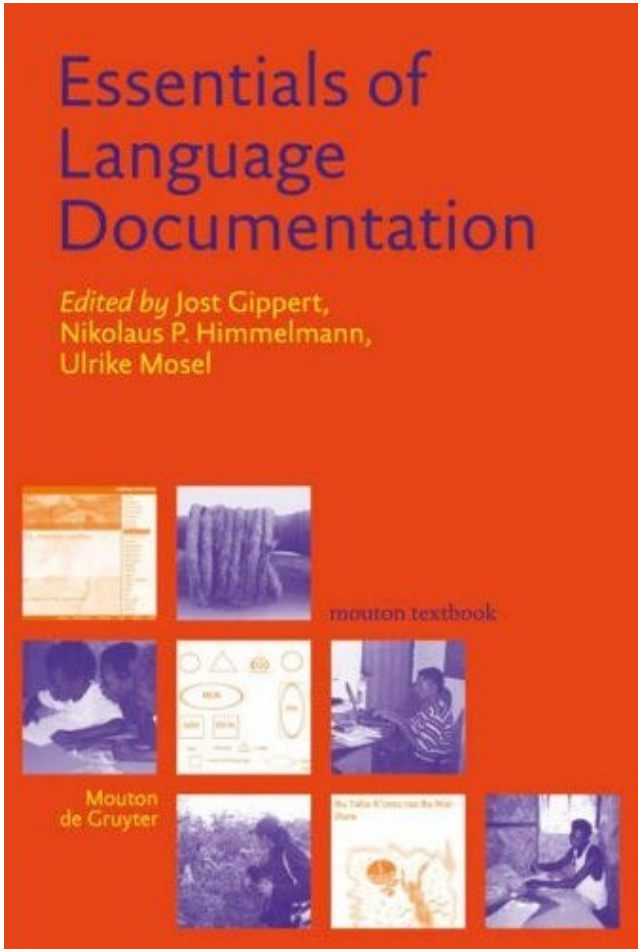
CLARIN

EUDAT

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

DASISH

2) Language documentation



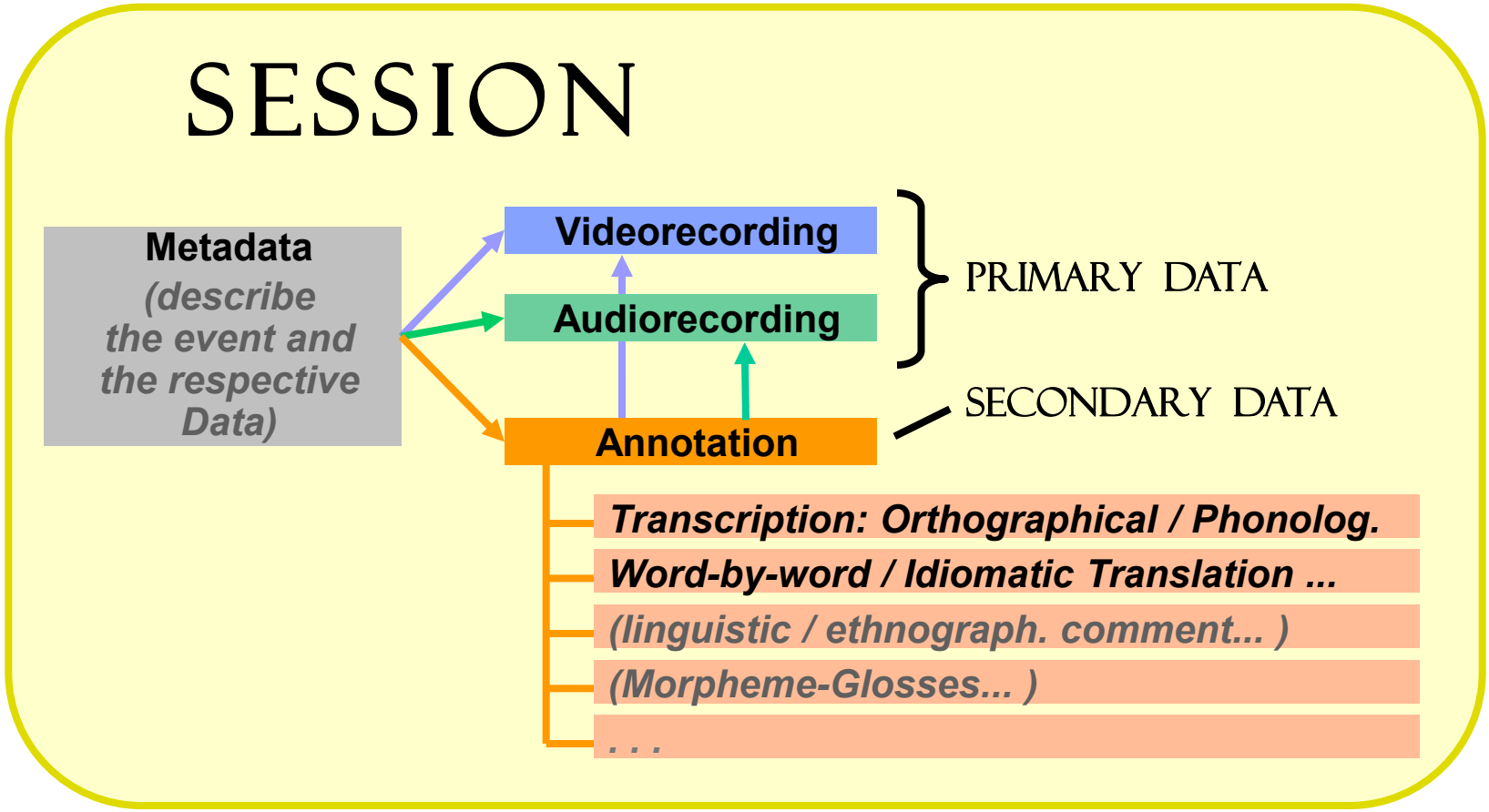
LANGUAGE DOCUMENTATION & CONSERVATION

rnld
Resource Network for Linguistic Diversity

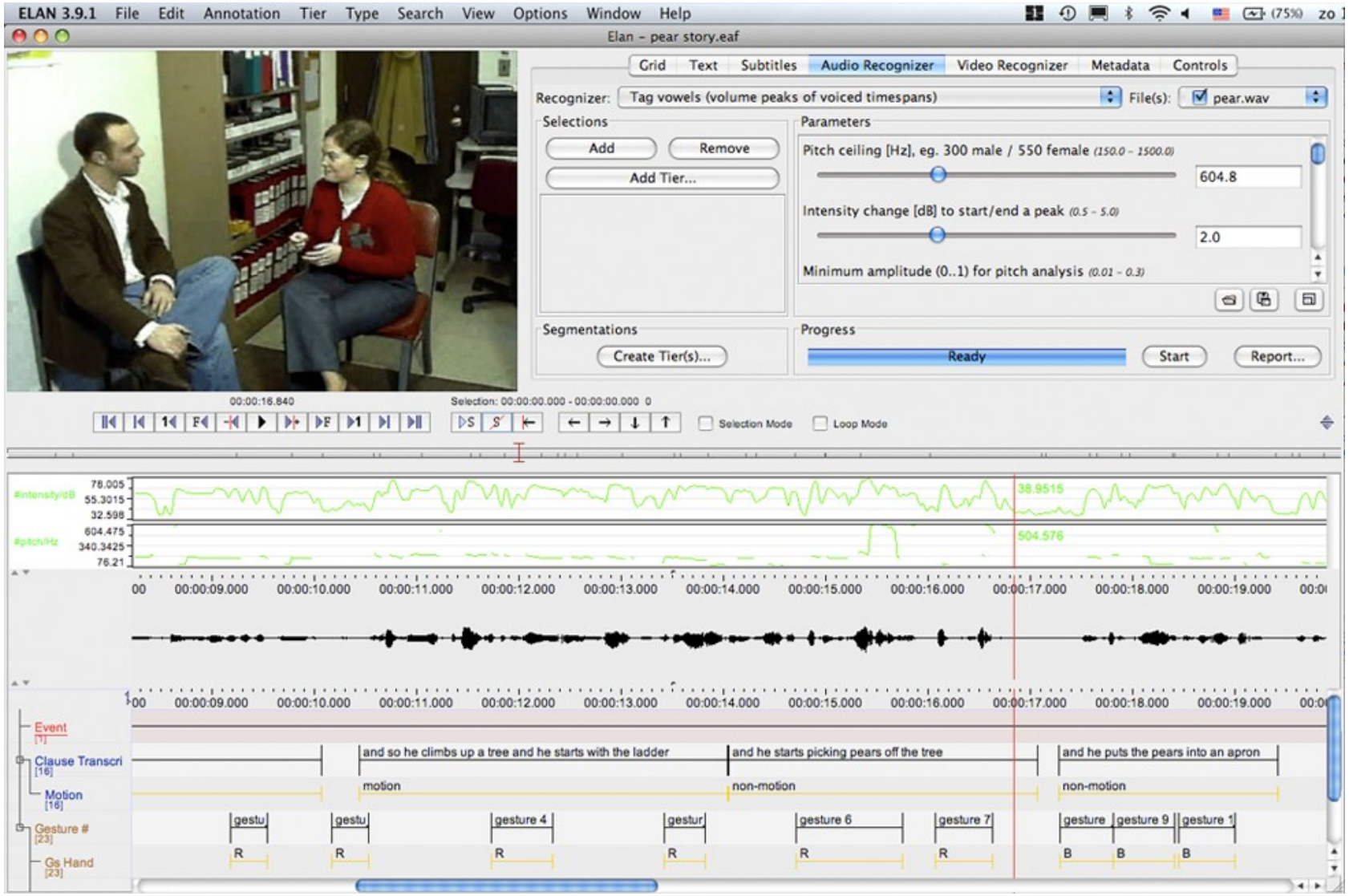
EMELD

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

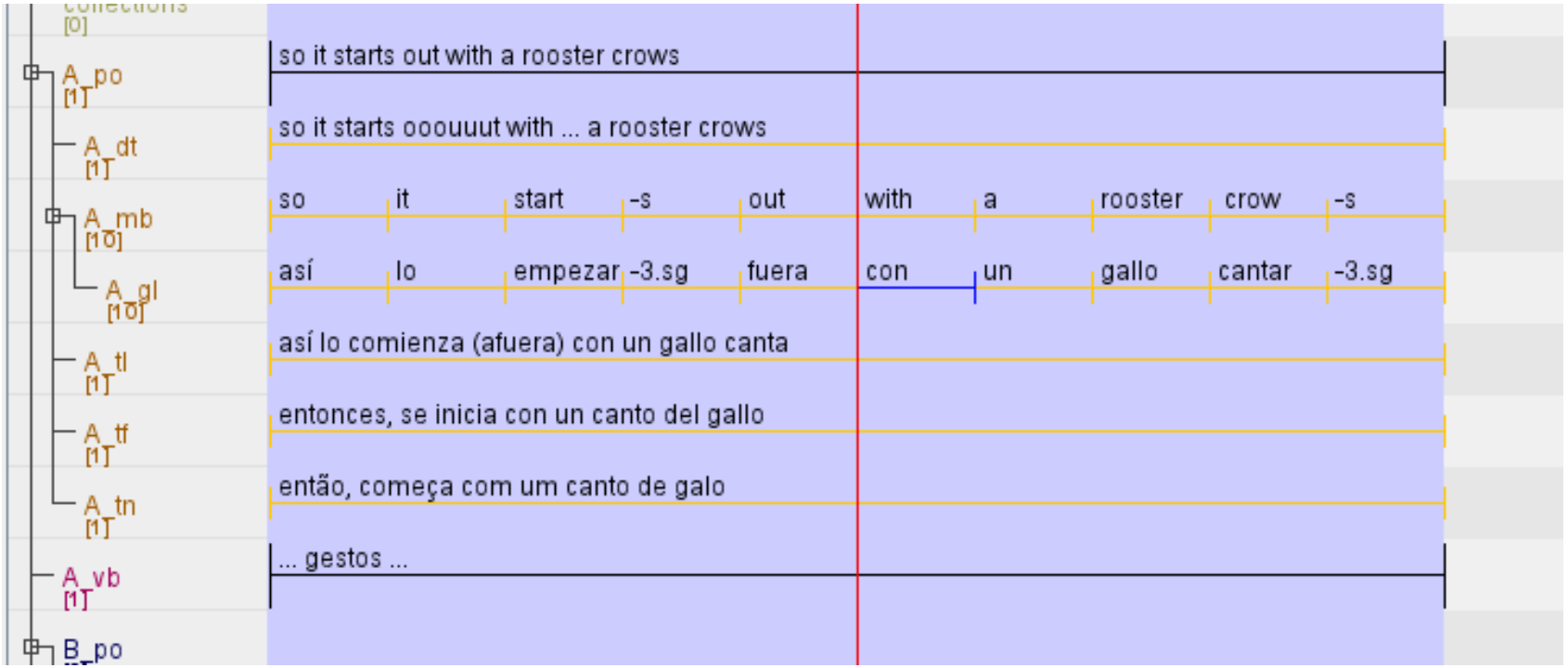
2) Language documentation



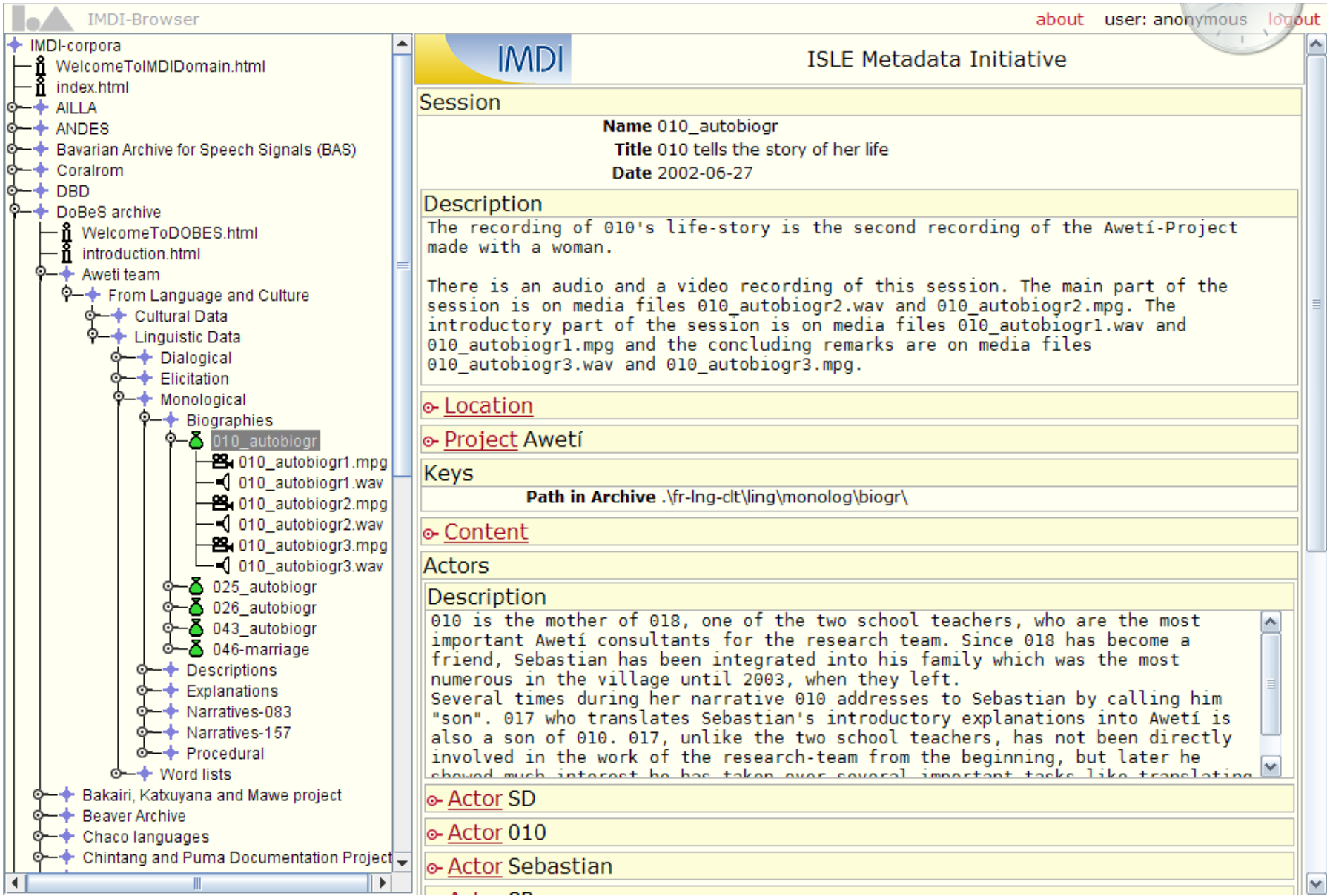
2) Language documentation



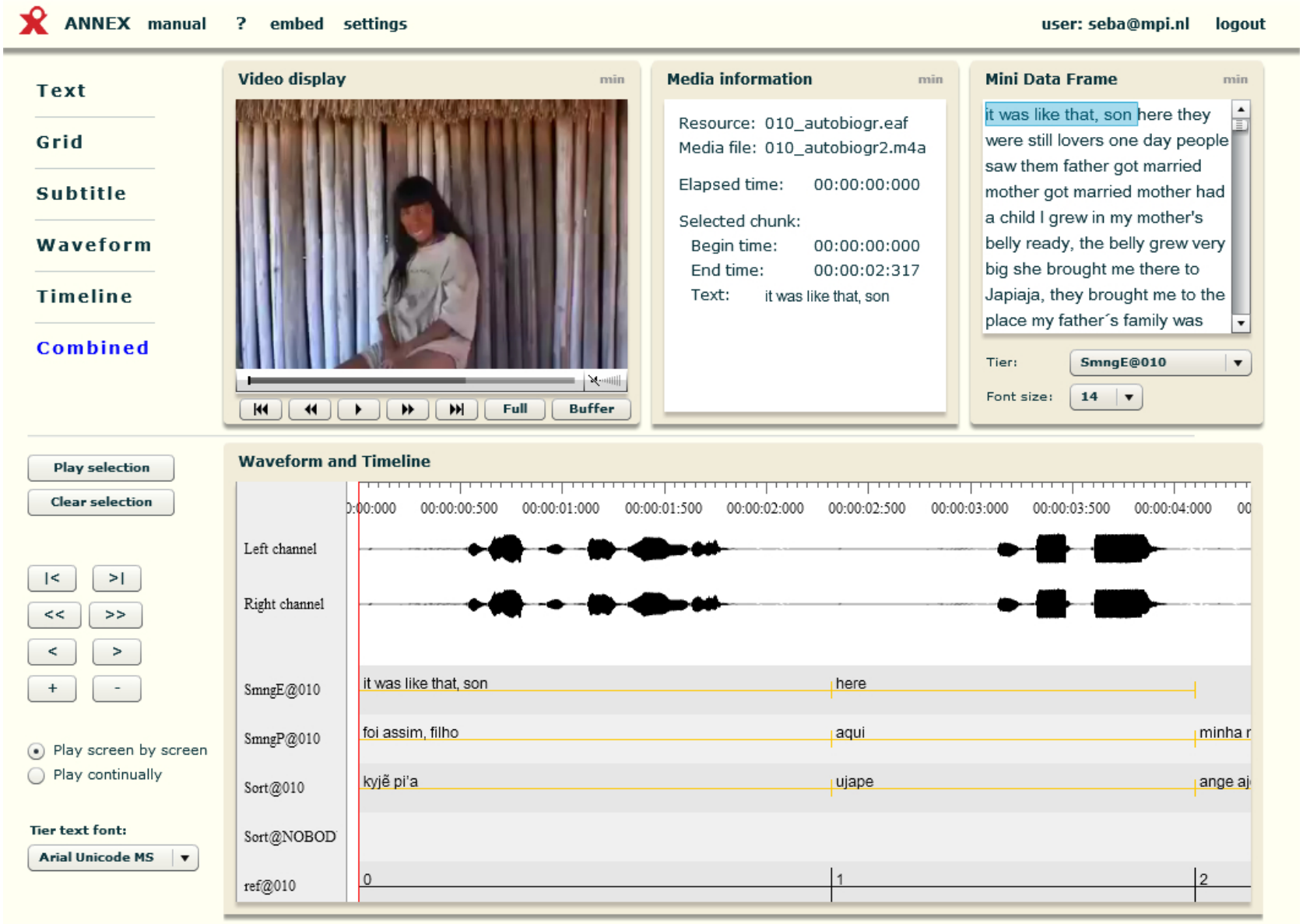
2) Language documentation



2) Language documentation



2) Language documentation



1) Language diversity and endangerment

2) Language Documentation

3) **The DOBES Programme**

4) The Language Archive

5) Relation to the Digital Humanities

3) The DOBES programme

General:

- Initiative by the VolkswagenStiftung together with German linguists
- A programme rather than a project: a self-organized research structure
- Independent research teams, steering committee, advisory boards
- The heart is one central technical project and archive at the MPI Nijmegen, now “TLA”
- Each project 2-4 years on one or more languages

3) The DOBES programme

History:

- Interest in topic “language endangerment” since 1991/92, LSA symposium etc.
- Summer-school in Köln in 1993
- Conceptional preparation (until) 1999
- Pilot Phase 2000/2001:
  5+3 individual projects & ‘TIDEL’ (TG @ MPI-PL)
- 2001–2012: calls for main phase,
  each year around 7 new projects
- Total of ca. 70 individual projects (28Mio €)
  on about 90 target languages

3) The DOBES programme

3) The DOBES programme

Achievements:

- General goals and methods have been established: that/how digital archives with primary data and annotation are built
- Relevant Software has been and is being developed (LAT: ELAN/ANNEX, LEXUS, IMDI/ARBIL, LAMUS/AMS)
- Legal and ethical questions have been adressed and partially answered
- National and regional archives are implemented
- Awarenes of (endangerment of) language diversity
- LD established as field, other funding programs

1) Language diversity and endangerment

2) Language Documentation

3) The DOBES Programme

4) The Language Archive

5) Relation to the Digital Humanities

4) The Language Archive

- New unit at the Max-Planck-Institute for Psycholinguistics in Nijmegen, Netherlands

4) The Language Archive

- New unit at the Max-Planck-Institute for Psycholinguistics in Nijmegen, Netherlands
- Director: **W. KLEIN**, Head: **PETER WITTENBURG**
- MPG, MPI-PL, BBAW, KNAW contribute
- Goal: give long-term stability for archive, software development and other activities
- One of the backbone centers of CLARIN
- Participation in many international projects:
- ISLE, MUMIS, ECHO, INTERA, DAM-LR, DoBeS, CGN, CLARIN EU, CLARIN NL, CLARIN D, CLARA, Inter, HARVE, AVATech, REPLIX, RELISH, *EUDAT, DASISH, INNET, iCORDI*

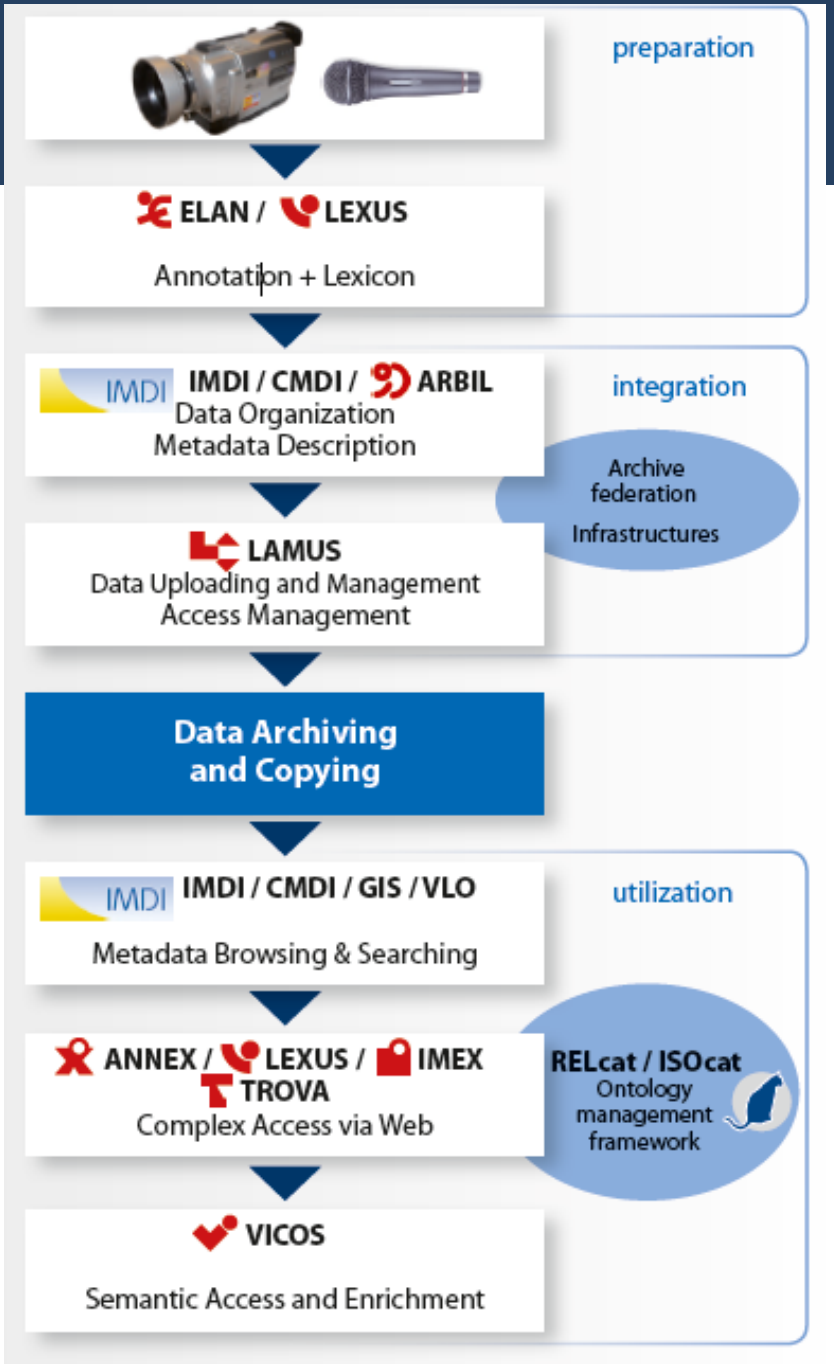
4) The Language Archive

State of Archive:

- about 60 Terabyte of well-described resources
  - Linguistic & ethnographical data from 200 languages
  - Language acquisition, sign language & multimodal dt.
  - Experimental, brain imaging, genetic data etc.
- about 20.000 hours of audio/video recordings
- about 73.000 metadata described “sessions”
- about 4.5 million annotated segments
- about 50 lexica
- DOBES: ca. 25TB on 65 languages

4) TLA

- Full lifecycle support
- New metadata format: CMDI (flexible, components can be chosen)
- ARBIL supports CMDI
- AVATecH: modules for automatic A/V recognition



Language Documentation & DH

1) Language diversity and endangerment

2) Language Documentation

3) The DOBES Programme

4) The Language Archive

5) Relation to the Digital Humanities

5) Relation to the Digital Humanities

- Diversity: languages, formats, approaches
- Sustainability: long-term availability, infrastructures (CLARIN etc.)
- Interdisciplinarity and teamwork
- Global network (archives, metadata)
- Open access, but also privacy rights and IPR
- Mobilization and explotation
- DH methods for reducing annotation costs

Intl' DH Congress 2012
2012-07-19
Hamburg

Language Documentation and Digital Humanities: The (DoBeS) Language Archive

Sebastian Drude, Paul Trilsbeek, Daan Broeder

The Language Archive – Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands