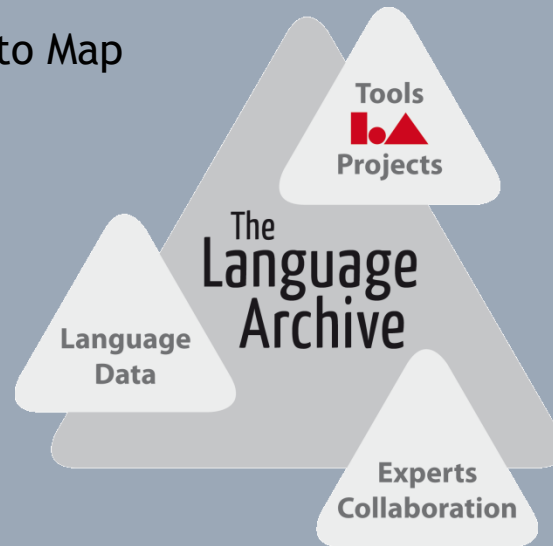


Charting Vanishing Voices:
A Collaborative Workshop to Map
Endangered Oral Cultures

World Oral Literature Project
2012 Workshop
CRASSH, Cambridge



MAX-PLANCK-GESellschaft

Sustainable Solutions for Endangered Languages Data: The Language Archive

Sebastian Drude, Daan Broeder, Paul Trilsbeek
The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands



- Introduction
- Sustainable data from linguistic fieldwork
- The Language Archive (TLA) @ MPI-PL
- Language Archiving Technology (LAT)
- Open access, legal & ethical issues
- Summing up: key challenges for sustainable data



- **Introduction**
- Sustainable data from linguistic fieldwork
- The Language Archive (TLA) @MPI-PL
- Language Archiving Technology (LAT)
- Open access, legal & ethical issues
- Summing up: key challenges for sustainable data



- Work of the Technical Group at the Max-Planck-Institute for Psycholinguistics in Nijmegen
- Now a new unit @ MPI: “The Language Archive”
- First as institutional solution for digital data (experiments, CHILDES, ESF-SL, fieldwork)
- From 2000 on: Central archive and technical centre of the DOBES programme (documentation of endangered languages)
- Integration with other centers and European data and research infrastructures



Introduction: DOBES



MAX-PLANCK-GESELLSCHAFT

- Initiative by the VolkswagenStiftung together with German linguists
- DGFS summer school 1993, first DOBES call 1999
- Independent research teams, steering committee, advisory boards
- The heart is one central technical project and archive at the MPI Nijmegen, now “TLA”
- Total of ca. 65 individual projects (28Mio €) on about 90 target languages
- Programme will end around 2016 (>15 years)



DOBES main features:

- Focus on data (linguistic analysis, revitalization and other activities welcome but additional)
- Language documentation in cultural context
- Interdisciplinary (e.g., Anthropology, Music-ethnology, Archaeology...)
- Partnership with community, training
- Emphasis on legal and ethical questions
- Common methodology and workflow
- Dissemination of language documentation (training courses, workshops, book)



- Introduction
- **Sustainable data from linguistic fieldwork**
- The Language Archive (TLA) @MPI-PL
- Language Archiving Technology (LAT)
- Open access, legal & ethical issues
- Summing up: key challenges for sustainable data



Sustainable data from linguistic fieldwork



MAX-PLANCK-GESELLSCHAFT

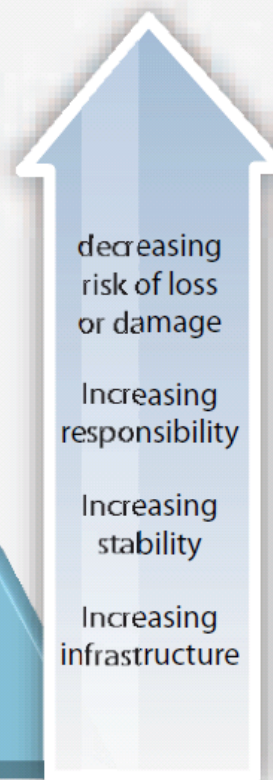
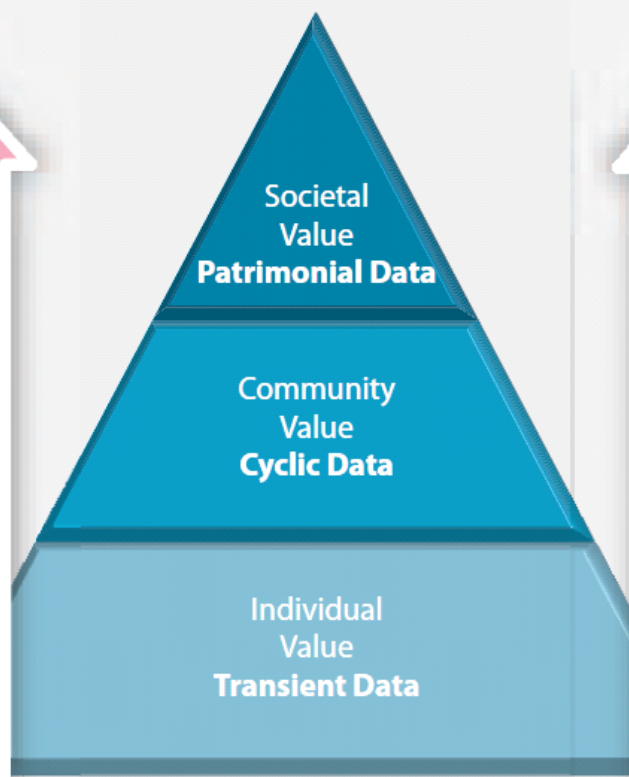
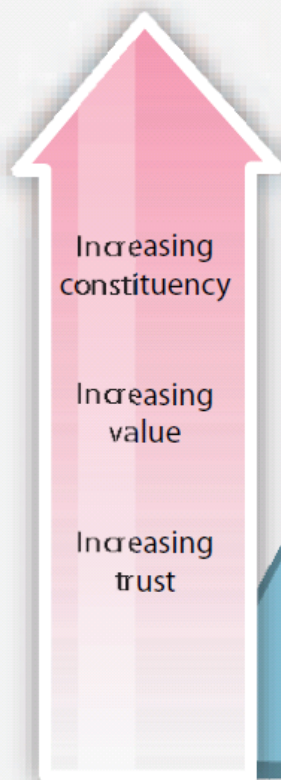
The data pyramid - a hierarchy of rising value and permanence

Digital Data Collections

Reference, nationally and internationally important, irreplaceable data collections

Key research and community data collections

Personal data collections



Repositories/ Facilities

National- and international-scale repositories, libraries, archives

“Regional” - scale libraries and targeted data archives and centers

Private repositories

Source: Adapted from Francine Berman, UC San Diego, in *Communications of the ACM*.

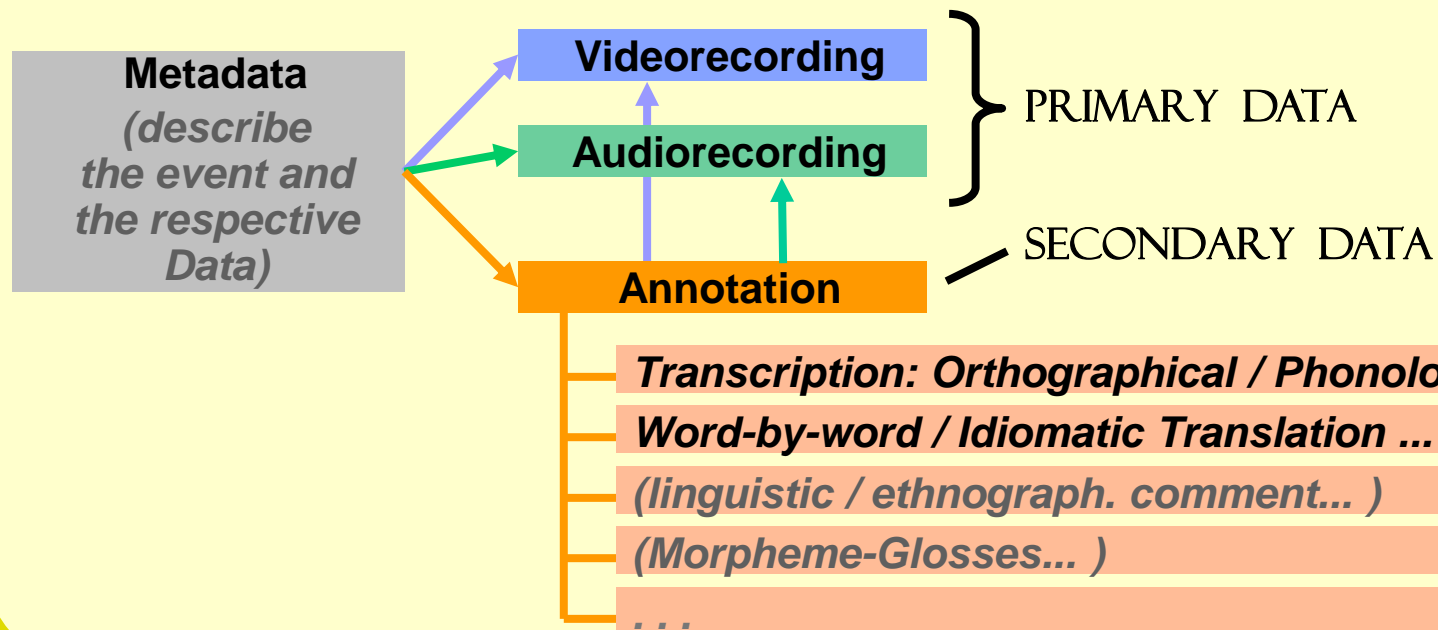


Sustainable data from linguistic fieldwork



MAX-PLANCK-GESELLSCHAFT

SESSION





Challenge of *sustainability*:

- ***Physical level***: limited lifetime of carriers
→ constant copying and replacement of carriers
- ***Logical level***: limited lifetime of formats
→ adherence to standards (Unicode, XML, open formats), constant updating of encodings
- Careful with transformations (lossy encodings, artefacts being introduced...), provenance info

Physical archives: “don’t touch!”

Digital archives: “touch frequently”



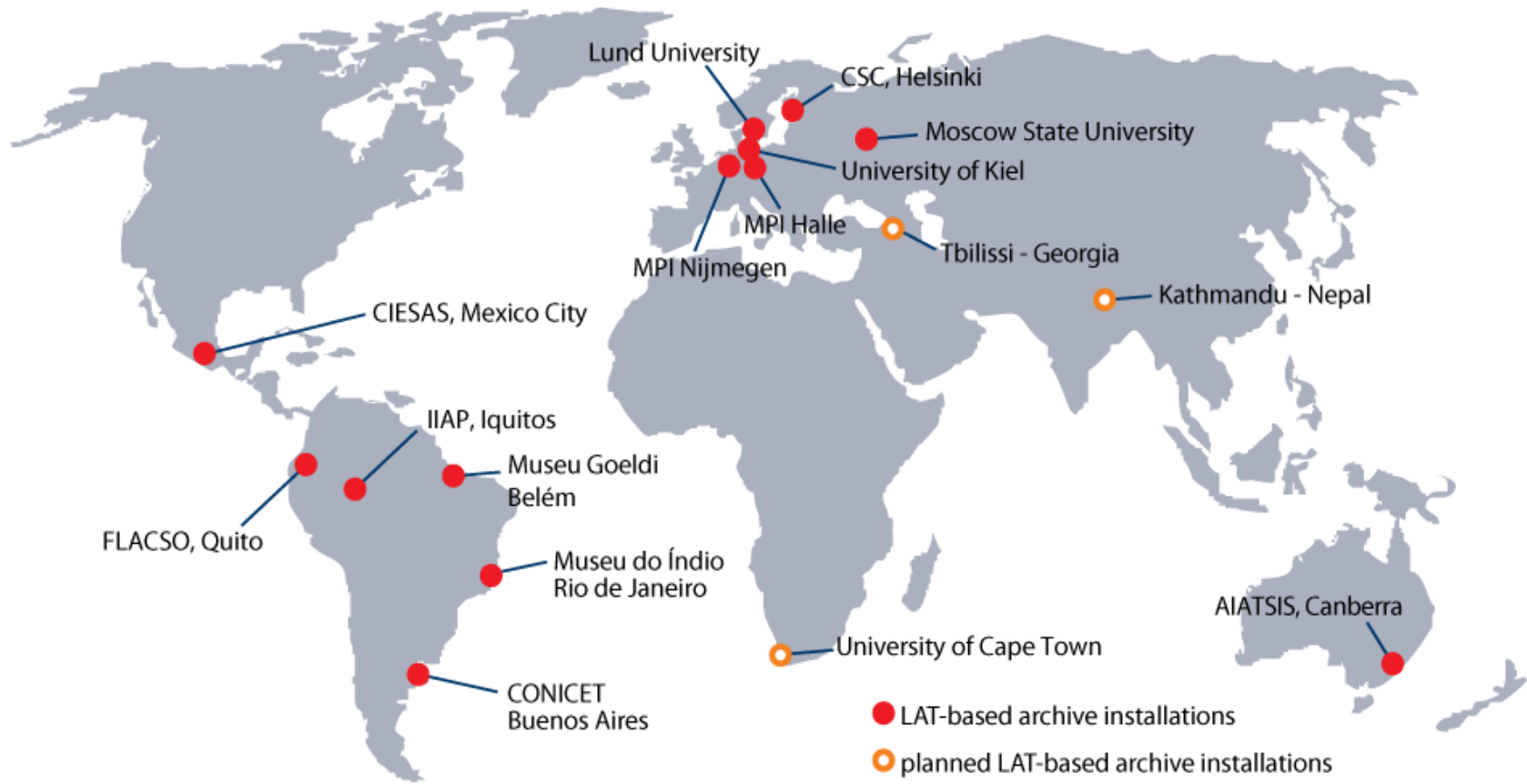
- Introduction
- Sustainable data from linguistic fieldwork
- **The Language Archive (TLA) @MPI-PL**
- Language Archiving Technology (LAT)
- Open access, legal & ethical issues
- Summing up: key challenges for sustainable data



- Currently: 80TB data in well-structured sessions
- PIDs (DOI / Handles) for all resources (versions), checksum... , implementing policy rules
- Data on ca. 200 languages, & CHILDES, Dutch...
- DOBES: 25TB on ca. 60 languages
- All data is online accessible (with access rights)
- Software and infrastructure development depends on project funding
- Establishing “The Language Archive” aims at a long-term perspective for a sustainable archive



“Regional” LAT archives





The Language Archive (TLA) @MPI-PL

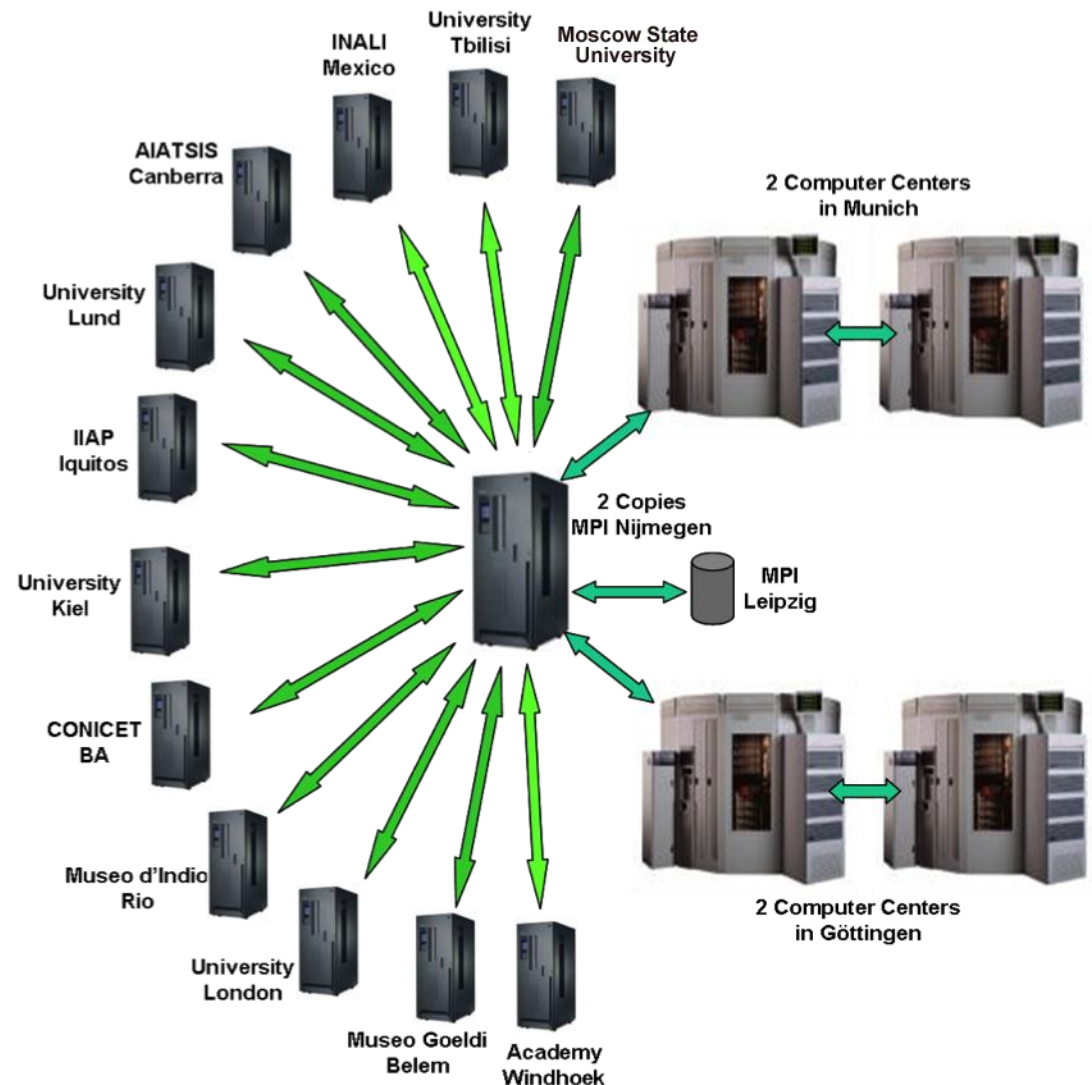


MAX-PLANCK-GESELLSCHAFT

Growing number
of archives using LAT

Six automatic full copies
at three locations
in Germany

Institutional guarantee
for bitstream-preser-
vation by the MPG
for 50 years





- Primary data: uncompressed PCM audio, MPEG video, in future jMPEG2000 (lossless compressed)
- Secondary data: Elan Annotation Format (XML-based, Unicode), “standard format” (Toolbox), and other open formats, also PDF
- Metadata: IMDI standard (in future: CMDI)
- Based on an integrated set of tools for archive administration and access, the “Language Archiving Technology” (LAT) suite of tools
- Regional archives based on LAT are being set up



Collaboration in larger projects (applying LAT):

- Leading role in different EU projects working on developing e-science infrastructure for the humanities (digital humanities / “eHumanities”)
- CLARIN (Common Language and Technology Research Infrastructure): Lang. Res. & Techn.
- DASISH (Data Service Infrastructure for the Social Sciences and Humanities)
- European Strategy Forum on Research Infrastructures (ESFRI)
- Cooperation outside Europe (DELAMAN, RELISH)

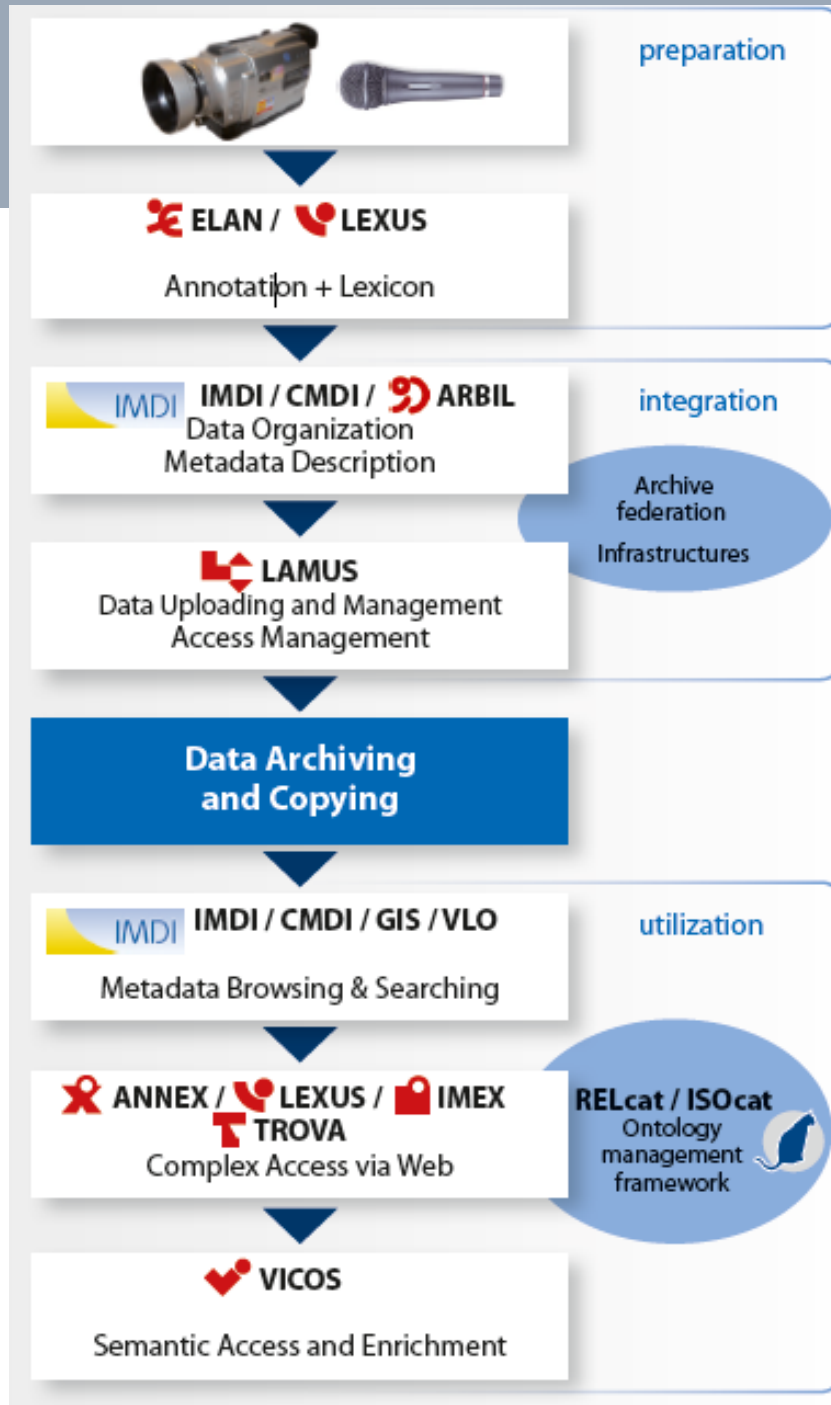


- Introduction
- Sustainable data from linguistic fieldwork
- The Language Archive (TLA) @MPI-PL
- **Language Archiving Technology (LAT)**
- Open access, legal & ethical issues
- Summing up: key challenges for sustainable data



Language Archiving Technology (LAT)

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





Language Archiving Technology (LAT)



MAX-PLANCK-GESELLSCHAFT

ANNEX manual ? embed settings user: seba@mpi.nl logout

Text

Grid


Subtitle

Waveform

Timeline

Combined

Video display min



⏪ ⏴ ⏵ ⏩ Full Buffer

Media information min

Resource: 026_autobiogr-2.eaf
Media file: 026_autobiogr2.m4a

Elapsed time: 00:00:12:900

Selected chunk:
Begin time: 00:00:04:484
End time: 00:00:08:220
Text: kype i'atu ti a'yn , kype me

Mini Data Frame min

(ite te) itemoajpu a'yt kype a'yn
itemoaipu a'yt kype a'yn kype i'atu ti
a'yn , kype me kype jatã net itã kitã
na'yt ne tazu'jyt tetam 'etu jatã pe a'yn
nanype a'ang a'yt i'atu a'yn kype mejka
itekozokotu a'yt kaminu'azan net
itekwahapu jatãtsu ika'akwahapejutu
mã 'jyt katu met akwawap wã a'yn
akwawap 'jyt jatã [mune] mã'jy[t]katu

Tier: Sort@026

Font size: 14

Play selection

Clear selection

|< >|

<< >>

< >

+ -

Timeline

00:08:500 00:00:09:000 00:00:09:500 00:00:10:000 00:00:10:500 00:00:11:000 00:00:11:500 00:00:12:000 00:00:12:500 00:00:13:000

SmngP@026 | isso se chama a aldeia da formiguinha

Sort@026 | jatã net itã kitã na'yt ne tazu'jyt tetam 'etu jatã pe a'yn

com@026

ref@026 | 4



Language Archiving Technology (LAT)



MAX-PLANCK-GESELLSCHAFT

The screenshot displays the Language Archiving Technology (LAT) interface, which is a web-based tool for linguistic research. It is shown within a Mozilla Firefox browser window. The interface is divided into several panels:

- Search Panel (Left):** Shows the search criteria: "Wichita corrected glottal stop". It includes a "List" of search results, with "hahri" selected. Below the list are navigation buttons for "First", "Previous", "Page 28", "Next", and "Last".
- Lexical Entry View (Middle-Left):** Displays the selected entry "hahri" with its meaning "vis" and "be angry". It lists several related entries:
 - 01 **neʔati:cháris** (he is angry) with examples like "neʔaʔ ta i2 uc hahri s2" and "bad pres pfocpat prevdat angry impf".
 - 02 **neʔataki:cháris** (I am angry) with examples like "neʔaʔ ta kiʔ uc hahri s2" and "bad pres oneob prevdat angry impf".
 - 03 **neʔah ke:ki:cháris** (I will be angry) with examples like "neʔaʔ keʔ kiʔ uc hahri s2" and "bad fut oneob prevdat angry impf".
 - 04 **neʔah ke:ki:cha:rʔi**
- Advanced Search Panel (Middle-Right):** Shows a "Galerie de photos" (photo gallery) with filters for "Lexicon" and "Filters". It includes a "Switch to structure view" button and a "filter" dropdown.
- Lexical Entries Panel (Bottom-Middle):** Lists various entries with their meanings and associated photo galleries:
 - 'apau kokomi: médicament; purge pour des bébés (medicine; purgative for babies) galerie de photo
 - faraoa: pain (bread) galerie de photo
 - haika 'eka kira: médicament; purge (purgative) médecine; purgative galerie de photo
 - haika 'epa: médicament; purge (femmes) médecine; purgative (women) galerie de photo
 - haika hati vaevae: médicament (jambe cassée) (medicine (broken leg)) galerie de photo
 - haika hó 'enana motua: huile de massage (massage oil) galerie de photo
 - haika hó toiki: huile de massage (massage oil) galerie de photo
 - haika ho'oi kivaiva: médicament (rein) (medicine (kidneys)) galerie de photo
 - haika kofé 'ehi: médicament (bébé, fille) (medicine (baby girl)) galerie de photo
 - haika mokio: médicament (bébé, fille) (medicine (baby girl)) galerie de photo
 - haika pūhenua: médicament (après accouchement) (medicine (after birth)) galerie de photo
 - haika putuhulu toiki: médicament contre bobos (medicine (spots)) galerie de photo
 - haika tekéo ika: médicament (empoisonnement) (medicine (fish poisoning)) galerie de photo
 - haika tōto me'ama: médicament (menstruation) (medicine (menstruation)) galerie de photo
- Photo Gallery Panel (Right):** Displays two photos related to the entry "'apau kokomi". The first photo shows a person's hand holding small, brown, cylindrical objects. The second photo shows a person's hand holding a small, white, cylindrical object. Both photos are labeled "apau kokomi" and "galerie de photo".



The screenshot shows the ISOcat web interface. On the left is a navigation tree with categories like 'My Workspace', 'Public', 'Thematic Views', 'Metadata', 'Morphosyntax', 'Basics', 'Cases', 'FormRelated', 'MorphologicalFeatu', 'Operations', 'PartOfSpeech', 'RegisterDatingFrec', 'Semantic Content Rep', 'Syntax', 'Language Resource C', 'Lexicography', 'Language Codes', 'Terminology', 'Multilingual Informati', 'Lexical Resources', 'Lexical Semantics', 'Translation', 'Sign language', 'Audio', 'CLARIN-NL', and 'GOLD'. The search bar at the top left contains the text 'intransitive'. The main area displays search results for 'intransitive' in a table format.

#	Name	Version	Administration	Registration st.	Char	Type	Owned by
1322	Intransitive	1:0	private	private	🚩	simple	Declerck, Thi
3080	AntiCausativeVoice	1:0	private	private	✓	simple	gold-user
3427	Processive	1:0	private	private	✓	simple	gold-user
3533	Transitivizer	1:0	private	private	✓	simple	gold-user
3457	Repetitive	1:0	private	private	✓	simple	gold-user
3247	ImpersonalPassiveVoice	1:0	private	private	✓	simple	gold-user
3276	Intransitivizer	1:0	private	private	✓	simple	gold-user
3548	Versive	1:0	private	private	✓	simple	gold-user
1225	absolute case	1:0	private	private	🚩	simple	Francopoulo,
3003	light verb	1:0	private	private	✓	simple	Francopoulo,

Below the table, a detailed view for the resource 'Intransitive - 1:0' is shown. It includes a '2. Description Section' with the following details:

- Profile: Syntax
- 2.1 Language Section**
- Language: English (en)
- 2.1.1 Name Section**
- Name: Intransitive
- Name Status: admitted name
- 2.1.2 Definition Section**
- Definition: Refers to a verb that does not take a direct object; that is, to a verb that does not express an action which directly affects another person or thing.
- Source: www.southwestern.edu/~carlg/Latin_Web/glossary.html
- 2.1.3 Example Section**



- Introduction
- Sustainable data from linguistic fieldwork
- The Language Archive (TLA) @MPI-PL
- Language Archiving Technology (LAT)
- **Open access, legal & ethical issues**
- **Summing up: key challenges for sustainable data**



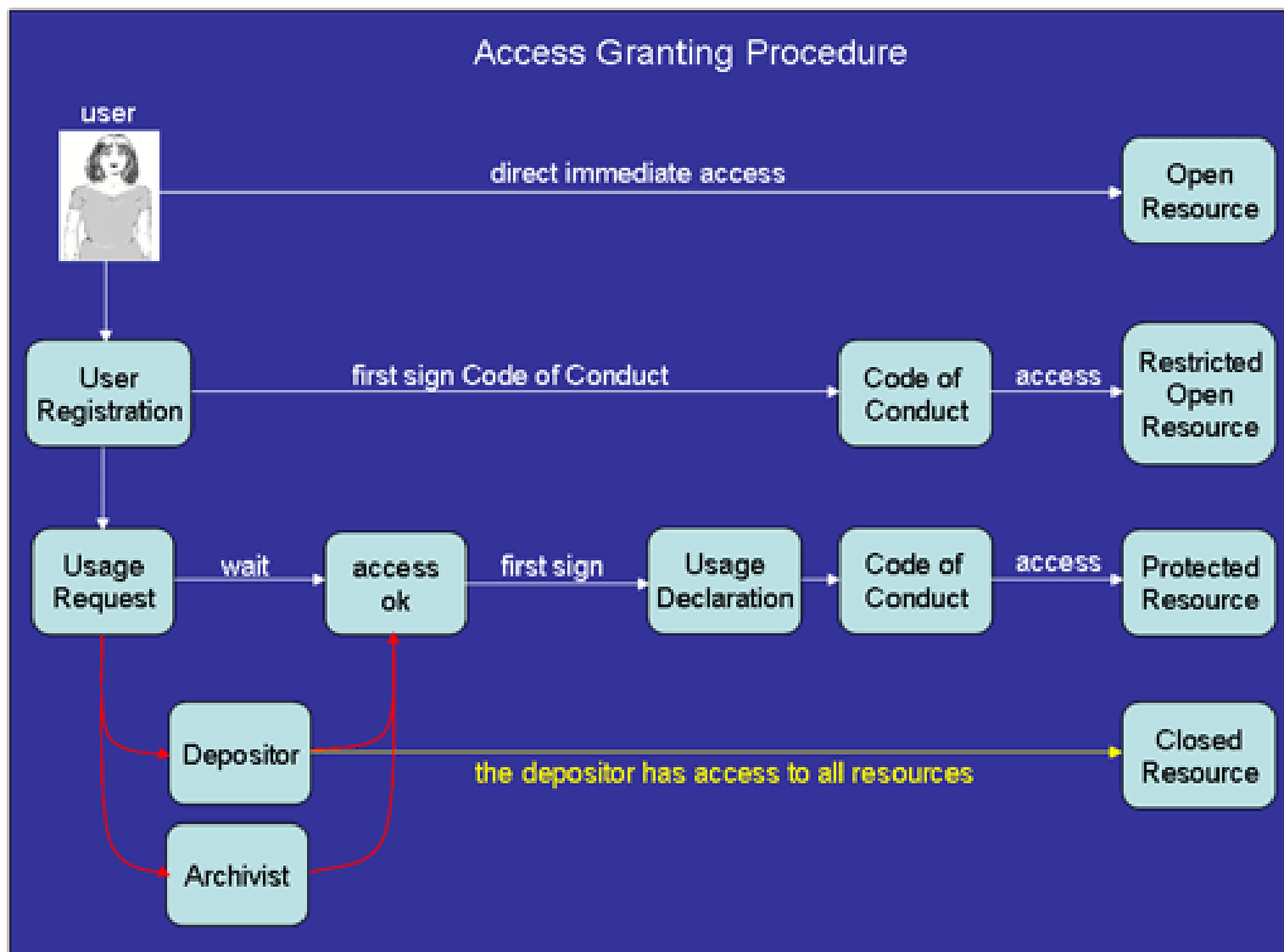
- Open access to research results *and data*
(Berlin declaration on Open Access to Scientific Knowledge)
- Accountability: EL data are irreproducible
- But: respect for privacy of human subjects
- Informed consent and anonymisation?
- Legal situation is complicated for all online-resources (national vs. international law etc.)
- DOBES: legal and ethical considerations are important (code of conduct, agreements, LAB)
- Trust between all parties is of key importance



Open access, legal & ethical issues



MAX-PLANCK-GESELLSCHAFT





- Introduction
- Sustainable data from linguistic fieldwork
- The Language Archive (TLA) @MPI-PL
- Language Archiving Technology (LAT)
- Open access, legal & ethical issues
- **Summing up: key challenges for sustainable data**



Summing up: sustainable data



MAX-PLANCK-GESELLSCHAFT

- Longevity:
 - (1) bit-stream → copies, migration,
 - (2) interpretability → standards, format update
- Access:
 - (1) identify & locate → metadata, search tools,
 - (2) retrieve & visualize content → access tools, download
- Public access: trust, Code of Conduct, responsibility, access management
- Provide data to the people we record: support for dedicated portals & enriched publications



Summing up: sustainable data

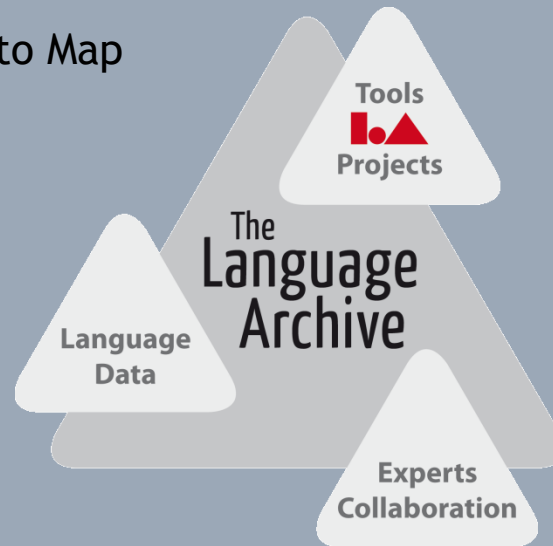


MAX-PLANCK-GESELLSCHAFT

- Maximum advantage of the access to data: A language archive is just one component in a digital research environment, interoperability
- Embed our analyses in accessible data: planned authoring environment for scholarly work
- Standards-conformant formats: good and useful tools will attract users (ARBIL, ELAN, in future with semi-automatic interlinearization function, possibly integrated with LEXUS, LMF, ISOcat)
- Support for most stages in the lifecycle of language documentation data

Charting Vanishing Voices:
A Collaborative Workshop to Map
Endangered Oral Cultures

World Oral Literature Project
2012 Workshop
CRASSH, Cambridge



MAX-PLANCK-GESellschaft

Sustainable Solutions for Endangered Languages Data: The Language Archive

Sebastian Drude, Daan Broeder, Paul Trilsbeek
The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands