The Language Archive

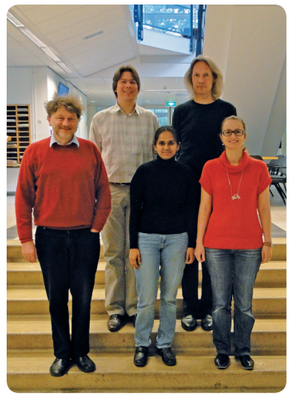
Language Archiving Technology LAT






State of Archive

- about 60 Terabyte of well-described resources
- about 20.000 hours of digitized audio/video recordings
- about 73.000 metadata described sessions
- about 4.5 million annotated segments
- about 47 lexica

The Language Archive

Storage and Access Systems

- production environment for all services:
  - 2 HP 8 core Xeon servers
  - Suse Linux Enterprise Edition
- storage system:
  - 2 Sunfire (4 core Opteron)
  - Hierarchical Storage Management System SAM-FS
  - 160 TB disk array
  - ADIC 2000 tape robot with 400 TB (extensible to PetaBytes)

Software

- all software in Java
- client technology in FLASH/ FLEX
- databases: PostgreSQL, HSQLDB, eXist, BaseX
- other software components: Shibboleth, Handle System, Lucene, OAI-PMH
- all own software is Open Source
- only usage of open source and free software components

Standards

- widely based on open standards
- archived resources preferentially make use of UNICODE; XML; generic models such as ISO LMF, ISOcat DCIF; RDF; XML-EAF; IMDI/CMDI; MPEG 2/1/4; mJPEG2000; JPEG/TIFF/ PNG; 48 kHz-16 bit linear PCM; OAI PMH
- regular quality assessment via Data Seal of Approval

The Language Archive

Max Planck Institute for Psycholinguistics

P.O. Box 310, 6500 AH Nijmegen
Wundtlaan 1, 6525 XD Nijmegen
The Netherlands

Phone: (+31) (0)24 - 352 19 11
Fax: (+31) (0)24 - 352 12 13
eMail: tla@mpi.nl

www.mpi.nl/tla

TLA is jointly funded by the Max-Planck-Society, the Berlin-Brandenburg Academy of Sciences and the Royal Netherlands Academy of Arts and Sciences

With substantial contributions by the VolkswagenFoundation, the European Commission, the German Ministry for Education and Research, the Dutch Science Foundation and the Max Planck Institute for Psycholinguistics.

MAX-PLANCK-GESELLSCHAFT

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN

Language Archiving Technology LAT

October 2011

TLA's Mission

- digitize and archive language resources
- support access to language resources
- develop tools, services and infrastructures
- set up of regional archives worldwide
- organize education and training activities
- give help and support

The Language Archive

Max Planck Institute for Psycholinguistics

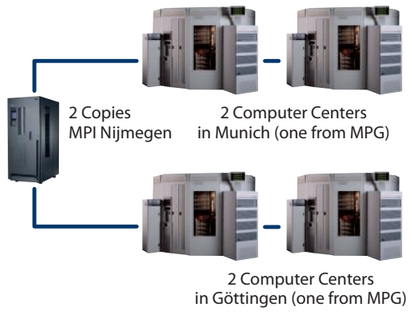
The Language Archive

DOKUMENTATION BEDROHTER SPRACHEN
DOCUMENTATION OF ENDANGERED LANGUAGES
DOBES

Language Archiving Technology LAT

Archive

The TLA builds on a large multimedia archive of language resources. To prevent the loss of valuable digital data, it is copied to various locations including a growing number of regional archives. To take care of the interpretability of data in the long run, adherence to standards and a continuous curation procedure are very important. Access to the data in the spirit of the ***Live Archives*** idea is guaranteed to those who have access permissions.



Long-term Preservation and Interpretability

- we face the loss of our cultural memory stored on electronic media
- UNESCO: 80% of the recordings about languages and cultures are highly endangered
- we need to preserve primary data (recordings, etc)
- we need to preserve secondary data (annotations, lexica, grammar descriptions, commentaries, metadata, etc)
- we need to preserve relations, context and provenance information
- TLA is committed to use open standards and to achieve high format coherence

Legal & Ethical Issues

- TLA requires the right to archive, but does not claim copyright
- the depositor decides on access permissions
- an agreed code of conduct is the basis for using the data
- 4 levels of access are foreseen from fully open to closed content
- access can be granted for a limited time

Collaborations

TLA is involved in a number of initiatives devoted to the archiving of digital language data, the improvement of technologies to create, manage and and access language data, and the construction of infrastructures that facilitate cross-institutional and cross-corpora access.

Deposits

An increasing number of researchers, in particular from the DOBES program, have been depositing resources of about 200 languages into the archive and are continuously enriching them further:

- about 60 DOBES teams are documenting more than 80 endangered languages.
- other researchers are depositing acquisition, speech, multimodal, multilingual, language and cognition, brain imaging, ethnological and other data

We will extend our activities in digitizing and archiving language data in order to make it available for research and to preserve it for the future.

Infrastructures

TLA participates in infrastructure projects on the integration and interoperability of data such as CLARIN (www.clarin.eu) and EUDAT (www.eudat.org). The resulting infrastructures will allow researchers to build virtual collections and workflows to improve data access in the direction of eHumanities usage scenarios. TLA also contributes to standardization in ISO TC37/SC4 (www.tc37sc4.org) which aims at facilitating interoperability in the language resources domain.

Technology Development

The LAT software is a set of components to create, manage, access, and enrich language resources. The major components will be maintained and extended. Future developments will focus on four pillars:

- use of statistical methods to automatically annotate textual, audio and video information
- extend existing tools to integrate them into powerful virtual research environments allowing to access a rich set of resources and tools in the Internet
- work on methods and tools that help to overcome the semantic interoperability problems
- offer tools as web services wherever possible and advantageous

Education and Outreach

Education and outreach activities need to be intensified to allow more researchers to participate in the eHumanities scenario and to preserve relevant language data. TLA will continue

- to set up regional archives at remote places
- to organize and participate in training courses at various places worldwide
- to organize and participate in summer schools to get young researchers involved.

Technology

The LAT software suite, started in 2000 with the multimedia annotation tool ELAN and the IMDI metadata infrastructure, covers about 15 components and tools. It is continuously being debugged, adapted and extended.

Resource Creation & Organization

ELAN, LEXUS, IMDI/CMDI, ARBIL, AV Recognizers

- create annotations and lexica
- create metadata and organize data
- support the use of standards
- support import and export options from/to tools such as Toolbox, CHAT, Transcriber

Management, Upload & Infrastructure

LAMUS, IMDI/CMDI, AMS, COSIX, HANDLE, REPLIX

- ensure archive consistency
- check uploaded formats
- create presentation formats
- create fast search indexes
- allow access rights definition
- add unique & persistent IDs
- basis is a robust storage and repository system with reliable mechanisms

Archiving Network

Regional LAT Archives have been set up in Canberra, Quito, Belem, Rio, Iquitos, Mexico City, Buenos Aires, Moscow, Lund, Kiel, Halle and 3 further setups are scheduled in 2012.

Projects

Past Projects: MUMIS, INTERA, ISLE, LIRICS, DAM-LR (all EC), CGN (NWO), HARVE, INTER, ROR (all MPG), REPLIX (DEISA, CLARIN)

Running Projects: DOBES (VWS), CLARIN (EU, NL, DE), DASISH, INNET, CLARA, EUDAT (all EC), AVATecH (MPG-FhG), RELISH (DFG/NEH)

Basic Resource Access

IMDI/CMDI/VLO

- metadata catalogue is essential for neutral access to single resources
- virtual collection building

Complex Access

ANNEX, IMEX, LEXUS, GIS, TROVA, VICOS

- access to annotated media, multimedia lexica and images
- content search on virtual collections
- creation of conceptual spaces and geographic overlays

Interoperability

ISOcat, RELcat, SCHEMAcat,

- ISOcat to register and define domain concepts
- RELcat to create, exchange and manipulate sets of relations between concepts
- SCHEMAcat to register and re-use schemas

Life Cycle Support

preparation

ELAN / LEXUS
Annotation + Lexicon

integration

IMDI / CMDI / ARBIL
Data Organization
Metadata Description

Archive federation
Infrastructures

LAMUS
Data Uploading and Management
Access Management

Data Archiving and Copying

utilization

IMDI / CMDI / GIS / VLO
Metadata Browsing & Searching

ANNEX / LEXUS / IMEX / TROVA
Complex Access via Web

RELcat / ISOcat
Ontology management framework

VICOS
Semantic Access and Enrichment