The Language Archiving Technology Domain

Alexander König
Jacquelijn Ringersma
Paul Trilsbeek

Max Planck Institute for Psycholinguistics
• only 10 years since we started
• 300,000 audio, video & text resources
• 100,000 metadata descriptions
• 19 TB of data

...and still growing daily
MPI researchers' data
DoBeS projects
Corpora from other research institutions
  • Corpus Nederlandse Gebarentaal
  • European Science Foundation's Second Language Acquisition by Adult Immigrants
  • Dutch Bilingual Database
The Three Pillars of Archive Management

The MPI Linguistic Archive

- LAMUS
- IMDI
- AMS
The Three Pillars of Archive Management

The ISLE Metadata Initiative Format

IMDI
• all data described using (XML-based) IMDI metadata format
• every resource/md description has a URID
• central organisational concept: the session resource bundle
  – obligatory metadata
  – a set of annotations & transcriptions
  – a set of media data (video, audio, photo)
IMDI – IMDI Browser

- all data in principle available online
  – http://corpus1.mpi.nl
• can be created/modified with IMDI-Editor
Soon Arbil will be released (testing version already available)
The Three Pillars of Archive Management

The Language Archive Management and Upload System

LAMUS
• Web-based operation
• Workspace principle
• Create / modify the (sub) corpus structure
• Upload / replace metadata & resources
• Gatekeeper function: not all data formats are accepted to keep archive consistent and future proof
• Assigning of URIDs
• Versioning system in the background
The Two Sides of Archive Access

ARCHIVE

metadata

Web apps.

Annotations media files

LAMUS

consumer
depositor
The LAMUS Interface

With the help of this page you can upload resources from your PC to the archive. Note that the uploaded resources will only be available to session-nodes that are linked directly from this node (or the parent node - as this is a session-node).

If you upload lmdi session files that have external entity definitions in a file "lmdi-sessions.lmdi" you must upload this file also!

You have already used 0 MB. There is 10240 MB available.

Use the "Browse" button in the graphical window below to select the resources going to be uploaded. This can be done repeatedly to select multiple files. You can also select whole directories. After you see all the required files in the window click the "Upload" Button of the graphical window to start the upload process.
The Three Pillars of Archive Management

The Access Management System

AMS
• Metadata is accessible for everyone (OAI)
• Resources are not always open for everyone
  • Sensitive information on actors
  • Ongoing research work with publication aims
• Access should be defined by researcher
• Different access rights for different users/user groups (access to video, audio and annotations can be set independently)
The AMS Interface

IMDI Browser - Mozilla Firefox

Focus overview on selected User/Group:

- - - EVERYBODY - - -

Rules of Node 'PeWi corpus':

- /MPI17/MP17/MP176399/MP176400 -

There are no licenses assigned to Node 'PeWi corpus'

- Manage Node Licenses

Rules of Node 'Demo':

- /MPI17/MP17/MP176399 -

There are no licenses assigned to Node 'Demo'

- Read Info Files Allow
  - Read Annotations Allow
  - Read Images Allow
  - Read Audio Files Allow
  - Read Video Files Allow
Accessing the Data

- TROVA
- IMDI Browser
- Metadata Search
- Google Earth Overlay
- Facetted Browsing
- LEXUS
- Web Portals
- ANNEX
- ViCoS
more on that in:
Eureka! User friendly access to the MPI linguistic data archive.
Accessing the Data - ANNEX

- ANNEX can be called from IMDI Browser
- Video with synchronised annotations
• Metadata is always open and can be searched by everybody
• If you have access to the resources, you can also search the annotations
ELAN
-
Time-aligned multimedia annotation
• supports interchange with a wide range of formats (e.g. Shoebox/Toolbox, CHAT)
• uses XML, Unicode (easy to convert from/to, future-proof)
• supports a wide range of media formats (playback is delegated to Media Player, QuickTime or JMF)
• cross-platform (runs on Windows, Linux and MacOS)
• Annotations on multiple tiers
• tiers can be hierarchically interconnected
• annotations can be directly time-aligned to media or refer to another annotation layer
• can link in up to four different video files
• can also link in both a video & an audio file
• highly configurable
• complex search options
• in constant development, incorporating user feedback
ELAN Interface
LEXUS
- Collaborative Lexicon Tool
LEXUS Functionalities

• web-based tool for creating & editing lexica
• workspace principle for multiple users
• complies to LMF ISO standard
• but still very flexible
• lexical entries can be enriched with multimedia (video, audio, photos)
• multimedia can simply be linked from the archive
**dada**

(N)

**fish sp (black spot sea perch)**

more than

Pintyö pum y:a dada doo ya,
more than 10 people were with Pintyö

should (tentative)
ViCoS
- Visualizing Conceptual Spaces
• supplement to LEXUS
• also web-based
• users can link concepts
• a number of “universal” relation types are provided
• users can define their own culture-specific relations
• idea is to bring indigenous people onboard
ViCoS Interface
Who manages the MPI linguistic archive

Archiving instance,
Max Planck Institute for Psycholinguistics

Archive managers: 3
Archive developers: 2
System manager: 1
Archiving software development: 4
Enrichment software development: 4
Who manages the MPI linguistic archive

Archiving instance,
Max Planck Institute for Psycholinguistics

Archive managers:
Alexander König
Jacquelijn Ringersma
Paul Trilsbeek
Who manages the MPI linguistic archive

Archiving instance,
Max Planck Institute for Psycholinguistics

**Developers:**
Daan Broeder, Marc Kemps-Snijders,
Han Sloetjes, Claus Zinn,
Peter Withers, Mariano Gardellini,
Thomas Koller
Thank you for your attention!