Large Multimedia Archive for World Languages

Przemek Lenkiewicz, Peter Wittenburg, Paul Trilsbeek
Max Planck Institute for Psycholinguistics

Motivation for the Archive

- Enormous speed of changes to languages and cultures, threat of losing a large part of our cultural heritage, knowledge about environment, species etc.
- There is an awareness about the need to document, archive and revitalize languages with Audio and Video recordings.
- Making recordings is not sufficient to guarantee that future generations will be able to access the data – standards and metadata are necessary.
- Platform for researchers to store and work on data.

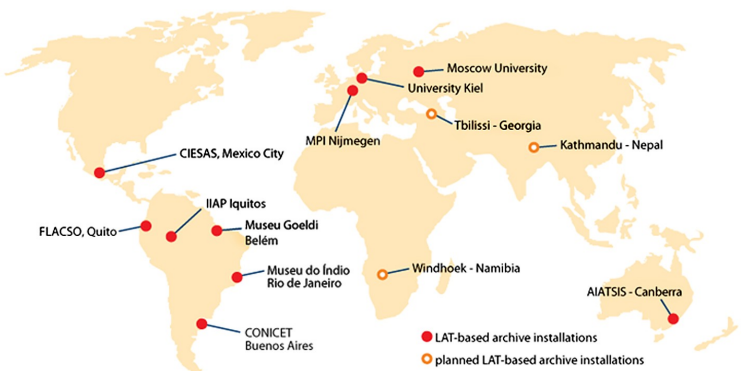
The Archive

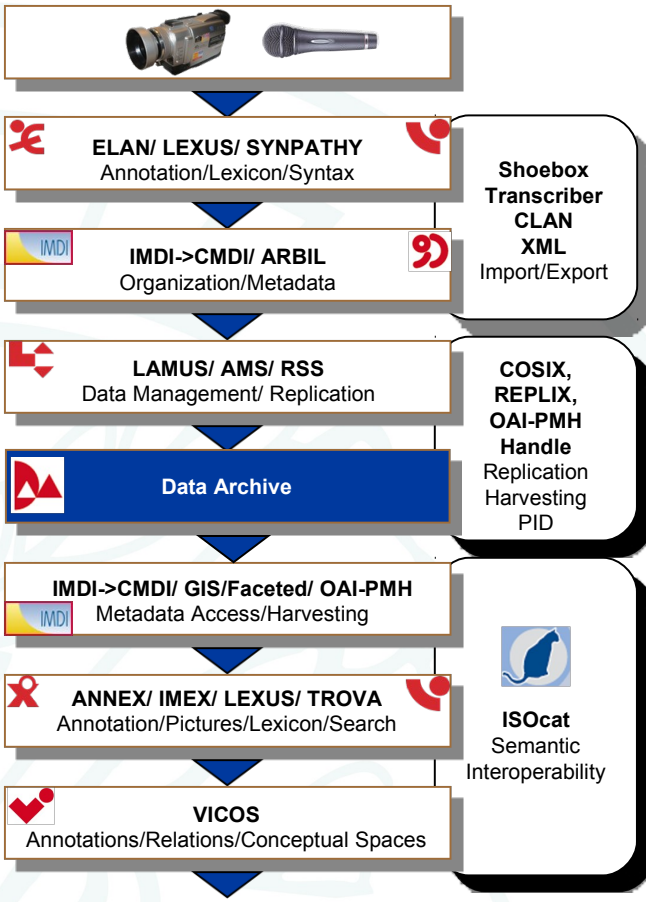
- Storage is based on two servers and the SAM-FS hierarchical storage management system with fast disk array, slow disk array and a tape library.
- At any upload immediately two copies are created.
- The archive currently stores more than 50 Terabyte of data contained in about 1 million objects.
- Two local copies are not enough: dynamic copies are created at two large centers in Germany, each of them having an agreement with another big computer center about long-term archiving.
- The president of the Max Planck Society has given a 50 years institutional guarantee for all resources stored at the computer centers.

Standards

- Very important for long-term interpretability is the adherence to standards where possible.
- A number of basic standards such as UNICODE and XML for texts, MPEGx for video, high quality linear Pulse Code Modulation (LPCM) for audio streams.
- No databases, no encapsulation allowed. All resources and metadata stored in the file system.

Network of Archives





Access To Archived Material

- Open source LAT (Language Archiving Technology) software suite, covers the whole lifecycle of language resource of different types:
- LAMUS and its components for access management and access requesting are acting as gate keepers for the archive to ensure consistency and coherence.
- ELAN and ANNEX for creating and viewing multimodal annotations for media recordings.
- ARBIL for combining metadata creation with organization capabilities.
- A number of web-applications have been developed to be able to access the archived material via the web: metadata browsing, searching, as overlay in Google Earth; content searching via TROVA engine.
- Thus the LAT software offers a comprehensive set of generic access technologies to users.

More information:
Alexander.Koenig@mpi.nl