

# Language Resource Archiving at the MPI

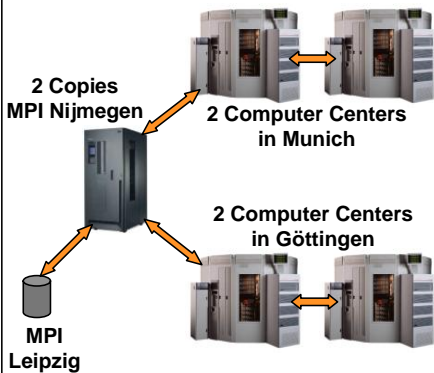
www.mpi.nl/corpus

## Language Resource Archives

- Many researchers want to upload their language resources into a well-organized repository.
- Long-term persistence can only be guaranteed by archivists.
- LR archives are modern archives that take care not only of persistence, but also of accessibility and enrichment.
- LR archives must separate the virtual organization that is relevant for users from the physical organization that is relevant for system managers. Due to new technology the physical structure will change transparent for the users.
- Archive managers and archiving software have to guarantee consistency.

## Long-term Persistence

- MPI dynamically creates currently seven copies at other data centers (Nijmegen (2), Munich (2), Göttingen (2), Leipzig(1))
- different protocols are used to cope with eventual errors
- MPSociety gives 50 years of institutional guarantee for its bit-streams
- interpretability is left to communities
- all 5 to 10 years migration to new technology
- working on world-wide distribution in DELAMAN



## MPI Vision

- the LR archive is the institutes capital
- it has to be organized, coherent and consistent
- LAMUS is the gate keeper to assure this
- it is the tool to upload and manage all resources
- it has an access management component
- the IMDI metadata infrastructure is used as basis
- unique resource identifiers guarantee persistence
- all is ready to connect with other LR archives
- all objects are accessible as individual resources
- ideally all objects are in a number of limited formats
- therefore everyone can write and use his own tools
- the archivist provides a number of access ways
  - for metadata browsing and searching
  - for utilizing simple objects (sound, video, texts)
  - for complex objects (lexica, annotated media)
- the archivist provides ways for archive enrichment

## LAMUS Language Management and Upload System

- web-based interface
- meant for users and managers
- define/upload archive structures, IMDI metadata and resources
- users have a workspace to play, arrange and test before commit
- creates indexes to be used by metadata and content search engines
- checks correctness of metadata
- configurable list of accepted file types, format parsers and complex types (lexica)
- parsers available XML, CHAT, EAF, LMF and Shoebox (to come)
- interaction with applications via API
- many checks included and to be included
- close relation with IMDI metadata tree in all respects
- access management system included
- introduction of unique resource identifiers to come

LAMUS is a complete Content Management System without encapsulating the files storing them in archival formats and allowing direct access.

LAMUS is easily portable to other institutes.

## State of Archive

- 11 TeraByte of data
- DOBES archive close to 1 TeraByte
- about 40.000 sessions, i.e. more than 100.000 objects
- included types: annotated media, lexica, grammars, field notes, ethnographic notes, ...
- included formats: EAF, XML, HTML, CHAT, Shoebox, PDF, WAV, MPEG2, MPEG1, MPEG4

