dLTP of Linguistic Resources at the MPI for Psycholinguistics

Paul Trilsbeek

MPI for Psycholinguistics &
DoBeS Archive

eScience Seminar Göttingen

19-20 June 2008

MPI digital language archive

- Started creating a digital archive about 10 years ago for a number of reasons:
  - Preventing “data cemeteries”: central, well organized data archive with rich metadata descriptions
  - Taking care of long-term preservation of data: continuous migration to latest storage technology, making use of “archivable” file formats, data distribution
  - Data access: Making data discoverable and accessible in various ways over the internet

MPI digital language archive

- Contains a large number of languages:
  - Languages from all over the world studied by MPI field linguists
  - First and second language acquisition studies
  - Endangered languages documented for the VolkswagenStiftung DoBeS programme
  - Spoken Dutch corpus
  - Sign languages
  - etc.
- Lots of unique material that cannot be recreated

MPI digital language archive

- Kinds of data:
  - video recordings
  - audio recordings
  - images
  - transcriptions/annotations
  - lexica
  - various other types of written documents
- All described using IMDI metadata scheme

IMDI metadata

- Developed in close collaboration with linguists to reflect their needs
- Used for both archive organization as well as resource description
- All IMDI metadata descriptions are XML files with links to resources. Databases are only used for fast searching and browsing.
- No “packages”, all resources and IMDI files are stored individually

MPI "LAT" Archiving Framework & Tools

Preparation

Primary resources

ELAN/LEXUS/SYNPATHY Annotation, Lexicon & Syntax

MPI "LAT" Archiving Framework & Tools

Preparation

Primary resources

ELAN/LEXUS/SYNPATHY Annotation, Lexicon & Syntax

Organization

IMDI: Metadata descriptions, data organization

LAMUS: Data Uploading and Management, Access Management

MPI "LAT" Archiving Framework & Tools

Preparation

Primary resources

ELAN/LEXUS/SYNPATHY Annotation, Lexicon & Syntax

Organization

IMDI: Metadata descriptions, data organization

LAMUS: Data Uploading and Management, Access Management

Archive

>35 TeraBytes
>10.000 hours A/V

MPI "LAT" Archiving Framework & Tools

Preparation

Primary resources

ELAN/LEXUS/SYNPATHY Annotation, Lexicon & Syntax

Organization

IMDI: Metadata descriptions, data organization

LAMUS: Data Uploading and Management, Access Management

Archive

>35 TeraBytes
>10.000 hours A/V

Utilization

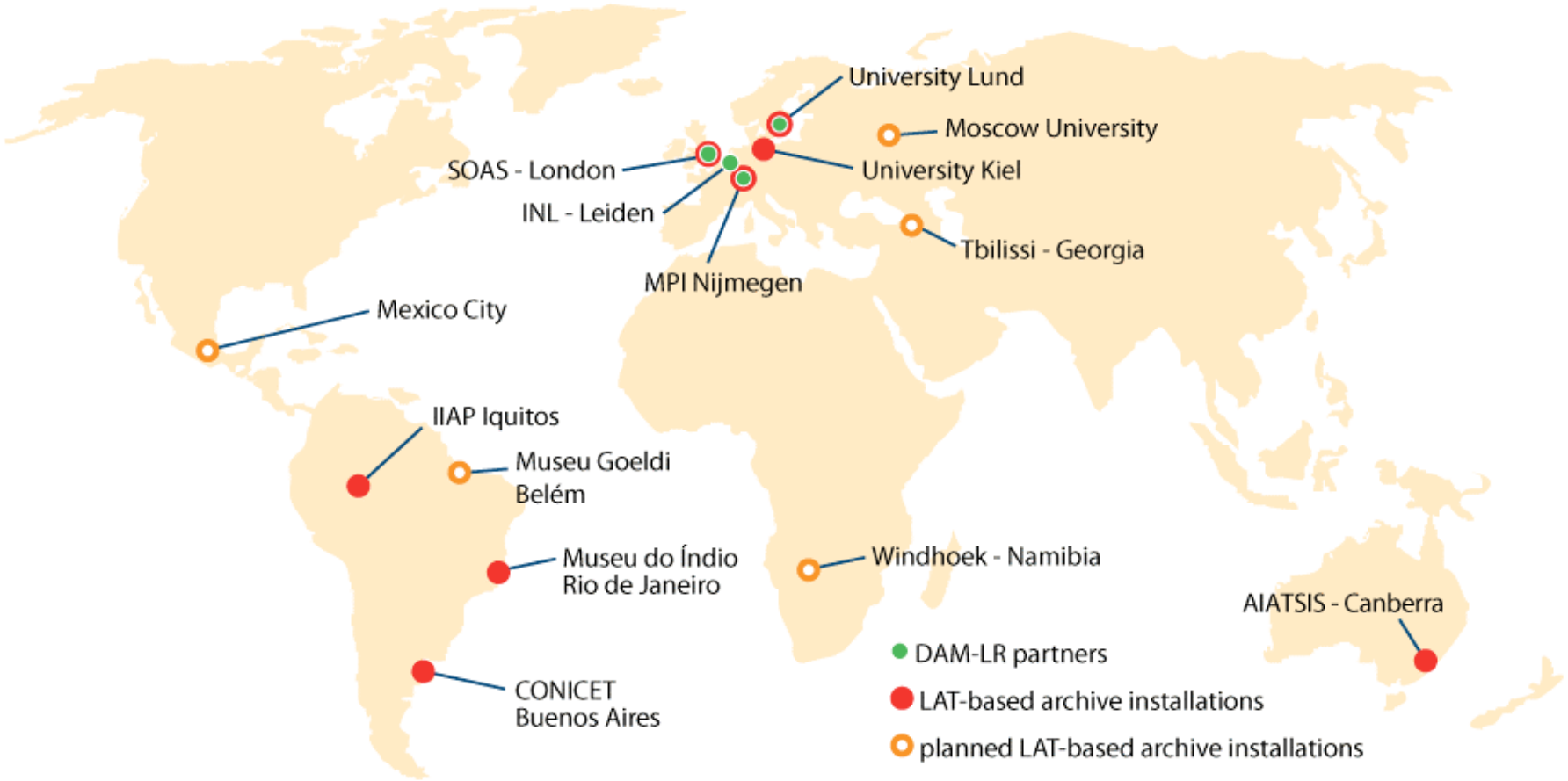
IMDI/GIS: Metadata Browsing & Searching, Google Earth

ANNEX/LEXUS/SEARCH: Complex Access via the Web

ADDIT/VICOS: Enrichments/Views

"Grid" of Language Archives

Distribution of MPI "LAT" archive installations



Long-term preservation

- Data Distribution:
  - 7 copies of each resource in various locations in the Netherlands and Germany
  - Guarantee from MPG president for

    bit-stream preservation for 50 years for our resources at GWDG and RZG
  - Grid of “Regional Archives” containing a part of the archived data that is relevant for that area

Long-term preservation

- Migration to newer storage technology every 4-5 years
  - Currently Sun SAM-FS HSM with ADIC Scalar i2000 LTO-3 tape library and

    2-level disc array
  - About 35 TB of archived resources at the moment
  - Migration to current system one year ago took about 2 weeks

Long-term preservation

- Minimize number of file formats and encodings
  - Linear PCM WAV for audio
  - MPEG2 for video (may switch to lossless MJPEG2000 in the future)
  - TIFF, JPEG and PNG for images
  - Various file formats for written material, if possible XML based and open
- Migration to newer standards should ideally be an automated process

Long-term preservation

- Persistent Identifiers
  - Making use of Handle PID system
  - Each resource (also previous versions) has its own persistent unique identifier
  - Gives stable references to resources for the long term

Collaboration

- DAM-LR: Distributed Access Management for Language Resources

  http://www.dam-lr.eu
- DELAMAN: Digital Endangered Languages and Music Archives Network

  http://www.delaman.org
- CLARIN: Common Language Resources and technology Infrastructure

  http://www.clarin.eu

Demo

- Archive browser & metadata search
- Content search & annotation viewer

New developments

- DANS “data seal of approval”

  http://www.datasealofapproval.org

  -> Audit of MPI archive will take place soon
- Revision of the metadata framework over the next few years within the CLARIN project

More Information

- www.lat-mpi.eu
- www.mpi.nl/dobes