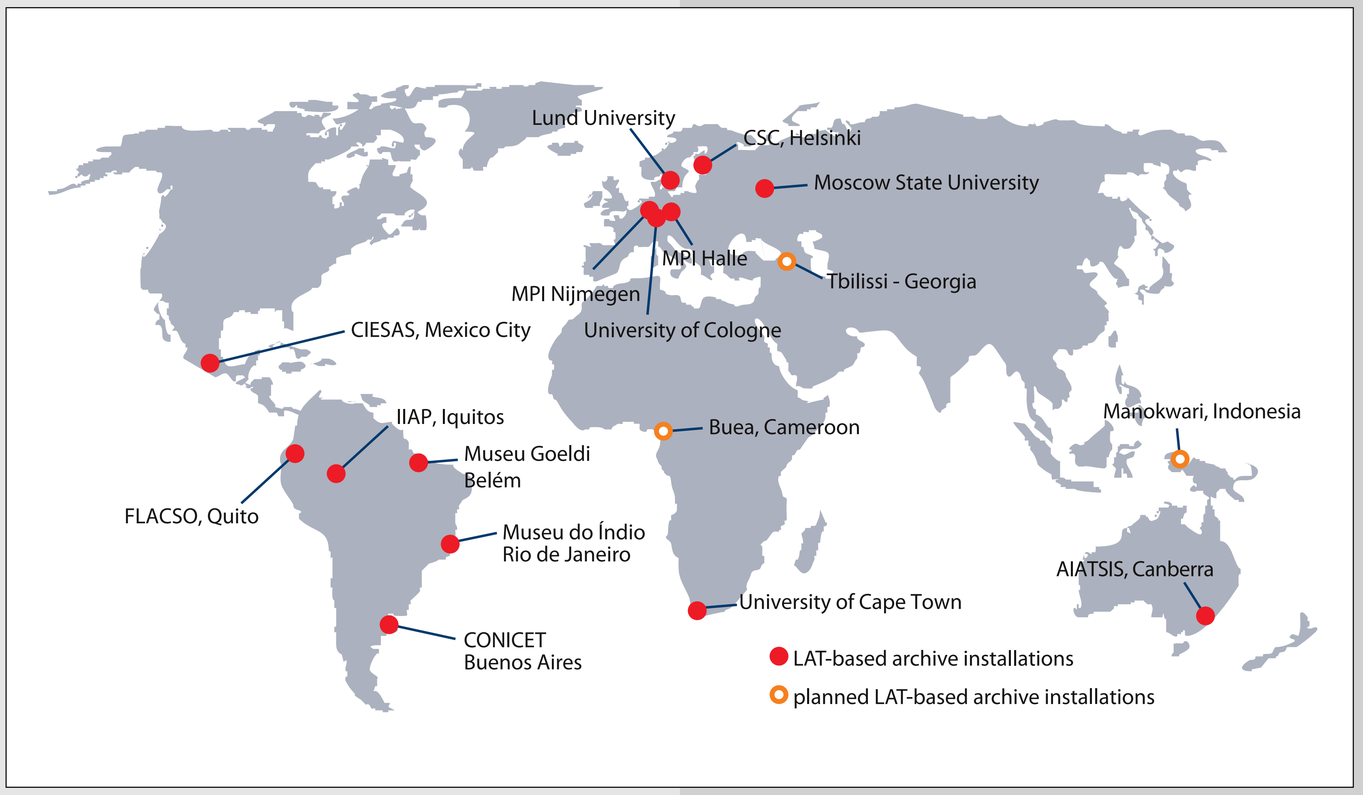
DOBES Archive

Long-term preservation and software development



Distribution of archives making use of the LAT archiving framework

The DOBES Archive

Archiving and long-term preservation of digital material is a task that requires specialised skills and is therefore best done by a dedicated institution, rather than by each researcher or research group individually. In the early phases of the DOBES programme, the DOBES Archive was established at the Max Planck Institute for Psycholinguistics in Nijmegen as the central place where all collected recordings and associated material from all documentation projects would be archived and preserved for the long term.

The DOBES programme played an important role in the development of the digital archiving framework that today forms the core of The Language Archive (TLA), a unit within the Max Planck Institute for Psycholinguistics that is concerned with the archiving of language data in general and the development of linguistic tools and language archiving software.

The Max Planck Institute for Psycholinguistics, where the DOBES Archive is located

Long-term preservation

Long-term preservation of digital material has two essential aspects:

- the preservation of the bits and bytes on digital storage media
- ensuring the interpretability of the files in the long run

Since all current data storage technology has a limited life expectancy, the so-called “bit-stream preservation” of data entails that one needs to migrate the data to new data carriers and up-to-date storage technology periodically, typically every 5 years at the moment. Bit-stream preservation also means that sufficient backup copies of the data should be created in different geographical locations, in case the archive would be destroyed due to a disaster such as fire or flood. Currently there are 7 copies of all data in the DOBES archive: 2 copies in Nijmegen, 2 in Göttingen, 2 in Munich (Garching) and 1 in Leipzig. In addition, some data is also copied to regional archives that are closer to the areas where the languages were documented and that make use of the same technological framework that is used for the DOBES Archive (LAT – Language Archiving Technology).

Keeping the data interpretable in the long run is also a challenge, since software programs and file formats typically do not have an eternal life either. Take for example the WordPerfect format, which was very common as a word processing format 15 years ago and is almost not used any more today. If you come across a WordPerfect file today, you will have a hard time opening it and looking at the content. An archive therefore needs to migrate files to newer formats before the current ones become obsolete.

A tape robot is used along with arrays of hard disks to store the data in the DOBES Archive.

What’s in the DOBES archive? (Feb. 2013)

- Total number of files: 124000
- Hours of video: 2500
- Hours of audio: 3700
- Written documents: 10800
- TeraBytes of data: 11

Tools and infrastructure development

TLA develops a range of tools for linguists in order to work with their recorded material and to organise and archive their collections. Many of these tools were developed in close collaboration with the DOBES documentation projects and with the help of DOBES funding. The table below gives an overview of the most important tools.

AMS	Tool for defining access permissions on archived material
Annex	Tool for viewing annotated video or audio recordings via the web
Arbil	Tool for describing and organising language data (metadata tool)
IMDI browser/ search	Tool for browsing and searching in the online catalog of the archive
ELAN	Tool for transcribing and annotating audio and video recordings
KinOath	Tool for creating kinship (e.g. family tree) diagrams and linking those to archived material
LAMUS	Tool for uploading and organizing data in the archive
LEXUS	Web-based tool for creating multimedia dictionaries (lexica)
Trova	Tool for searching in transcriptions and annotations in the archive
Vicos	Tool for visualising conceptual relations between entries in a LEXUS dictionary

TLA is also heavily involved in the development of large-scale research infrastructures for linguistics and for the humanities in general. The projects in which these infrastructures are developed, such as CLARIN and DASISH, aim at interconnecting the currently fragmented landscape of services and tools that are available to the various research communities.

The DOBES Archive forms an interesting use case for these infrastructure projects due to unique irreplaceable nature of the recordings that are archived and the sensitivity with respect to providing access to some of this material.

CONTACT ADDRESS

Paul Trilsbeek
DOBES Archive / The Language Archive
Max Planck Institute for Psycholinguistics
P.O. Box 310
6500 AH Nijmegen
The Netherlands

Paul.Trilsbeek@mpi.nl
http://www.mpi.nl/DOBES
http://tla.mpi.nl

VolkswagenStiftung