
# A sustainable archiving software solution for The Language Archive

Paul Trilsbeek, Daan Broeder,
Willem Elbers, André Moreira

The Language Archive
Max Planck Institute for Psycholinguistics

# Outline

The Language Archive

- History of TLA and its archiving solution
- Reasons for looking for a new solution
- Requirements
- Possible solutions
- Choices and developments so far

# History of TLA

The Language Archive

- Started towards the end of the nineties as part of the Technical Group of the Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands
- Archiving of audiovisual language corpora, development of linguistic tools (e.g. ELAN) and development of language archiving infrastructure
- In 2000 became the central archive for the DOBES language documentation programme funded by the Volkswagen Foundation

# History of TLA

The Language Archive

- Numerous external projects contributed to the development of tools and infrastructure (e.g. ISLE, DAM-LR, CLARIN, EUDAT)
- In 2011, TLA established as a separate unit within the MPI, core positions funded by Max Planck Society, the Berlin-Brandenburg Academy of Sciences and Humanities, and the Royal Netherlands Academy of Arts and Sciences
- Around 30 staff members in 2012

# History of TLA

The Language Archive

- In 2014, MPI directorate decided that TLA in its present size and form could not be supported in the long run and proposed a reorganisation
- Effective from October 2016:
  - Archive and some development of archiving and linguistic tools to stay in Nijmegen
  - General infrastructure projects and developments to be done through the Computing Centre of the Max Planck Society in Garching (Munich)

# Why a new archiving system?

The Language Archive

- Fewer software developers available in the long run
- The current system has been developed completely in house over a period of almost 15 years. It is rather complex and therefore costly to install and maintain
- Meanwhile, repository solutions have been developed that are actively maintained by open source communities (as well as commercial companies) and are widely used by archives and research data repositories around the world

# Requirements

The Language Archive

- The new system should have many of the features of the current system, in particular
  - Being able to quickly search and browse through hundreds of language corpora containing more than half a million files in total
  - Being able to search and browse in CMDI metadata records and visualise those
  - Being able to define various levels of access to certain users and groups
  - Offer an easy to use self-deposit web interface such that depositors themselves can upload and organize their data and metadata

# Requirements

The Language Archive

- Visualisation of audio-visual media
- Visualisation of annotated media (possibly by re-using the current ANNEX component)
- Offer a way to search the content of annotations and other textual resources (possibly by re-using the current TROVA component)

# Requirements

The Language Archive

- Solution should be based on well maintained and widely used open source software as much as possible
- Solution should ideally use programming languages and frameworks of which expertise is present within TLA

# Solutions

The Language Archive

- Repository solutions that were considered:
  - DSpace
  - Fedora Commons
  - EPrints
  - Greenstone

# Solutions

The Language Archive

- Repository back end: 2 serious candidates
  - Fedora Commons
  - DSpace
- In case of Fedora Commons: 2 front end candidates:
  - Islandora
  - Project Hydra

# Solutions: Back end

The Language Archive

- Fedora Commons:
  - Rather flexible in accommodating almost any kind of data model
  - No turnkey repository, more a framework to build a repository with. (However, in combination with Islandora front-end can be used pretty much out of the box)
- DSpace:
  - Turnkey repository, comes with front-end.
  - More limited data model support
  - Limited possibilities for modifying front-end without substantial changes to the codebase

# Solutions: Back end

The Language Archive

- Fedora Commons chosen as the better option for TLA given the flexibility of data models and the desire for customizing the front-end to suit the needs of a language archive

# Solutions: Front End

The Language Archive

- Front end options: Islandora
  - Drupal CMS on top of Fedora
  - Written in PHP
  - Modular setup
  - Allows the use of Drupal features and add-on modules as well
  - Drupal not that fast for large amounts of content
  - Difficult upgrading between major Drupal versions
  - Deposit interface not really suited for self-deposit by researchers

# Solutions: Front End

The Language Archive

- Project Hydra
  - Seemingly rapid development cycles
  - Hydra-based easy to use self-deposit solutions exist, however only for "simple use case" institutional repository
  - Not really meant for out of the box deployment but rather for building your own solution (except for “simple use case” institutional repository solutions)
  - Ruby on Rails framework

# Choices

The Language Archive

- Pursuing Fedora Commons / Islandora combination now
- Will use built in features for browsing, searching, and visualisation as much as possible
- Extensions / additions as much as possible in the form of Drupal modules (Islandora solution packs)

# Challenges

The Language Archive

- Easy to use deposit tool
- Flexible access permissions system
- Dealing with substantial changes in the chosen frameworks
  - Fedora Commons 4 (released Dec. 2014)
  - Drupal 8 (2015?)

# Roadmap for "EasyLAT"

The Language Archive

- Version 1 (Feb. 2015): all data and metadata of TLA ingested in Fedora/Islandora instance, browsable and accessible
- Version 2 (Oct. 2015): CMDI metadata editing, searching and visualisation, access permissions can be defined, user friendly data deposit feature
- Version 3 (June 2016): text/annotation content search
- Production ready Oct. 2016 at the latest