Ingest

Ingest into The Language Archive is done with the LAMUS tool (https://tla.mpi.nl/tools/tla-tools/lamus/). With this tool, the depositor (researcher or archivist) selects a (sub-)collection from the repository to which she/he will add materials. Initially the collection is empty but for subsequent deposits, a copy of the collection is made into a workspace. The depositor adds materials (data and metadata) to the workspace, creates a hierarchical structure within the workspace, and relates the newly deposited materials to this structure. Uploaded materials are immediately validated. Invalid data or metadata as well as data formats that are not accepted in the repository are rejected from the workspace.

Once the depositor is done with adding materials to the workspace and is satisfied with the changes, he/she can submit the workspace for inclusion into the repository. During this step, the actual ingest will take place after another verification step and assignment of Handle persistent identifiers for each object. Changed or replaced objects will be versioned, objects that stay the same are not touched in the repository.
Archival Storage and Data Management

Archival storage takes place on an Oracle SAM-QFS Hierarchical Storage Manager system. This system is set up such that automatically two backup copies of each archived object are created on LTO tape. This system also maintains md5 checksums for each object, which are used to verify integrity of the various copies. In addition to the local copies, 4 more backup copies are created in 4 different locations in Germany (2 institutions in Göttingen and 2 institutions in Garching near Munich).

Repository consistency and availability are monitored automatically using scripts and monitoring tools. Scripts are used to validate whether the information in the archive system’s Postgres databases is consistent with the data on the archival storage. Nagios monitoring software is used to continuously monitor the availability of web-based services. Munin monitoring software is used to continuously monitor all relevant components of the repository hardware.

Disk storage is in the form of RAID disk arrays and is typically replaced every 4 to 5 years. Individual disks are replaced earlier when failures occur, standard RAID functionality is used to rebuild the missing data. LTO tape storage drives and tapes are typically replaced every 5 years. Data integrity of files on tape is verified upon each access, tapes are replaced in case of errors and new copies on tape are created from copies that are still intact (on disk or tape).
Repository Access

Access to resources in the repository is provided by means of custom built web-based browse and search tools. All data sets within the repository are organised in a hierarchical structure, which can be browsed online. Metadata can be searched by means of a simple search box to search across all metadata fields, or an advanced search interface that allows the selection of one or more specific metadata fields. Textual resources are indexed and an advanced search interface is provided to enable linguistically relevant methods of searching within texts.

Direct access to all individual archived files is provided, either via the browse and search tools, or via the Handle persistent identifier which links to a unique URL per file. For audio and video materials, lower bandwidth access copies are generated automatically and made available in addition to the original objects.

For transcribed and annotated audio-visual materials, an advanced web-based viewer is available.

All metadata are openly accessible; data objects may be subject to access restrictions depending on the judgement of the depositor regarding the appropriateness of public access to the objects.