The Language Archive

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
True - accept

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
accept

Brief Description of Repository

The Language Archive (TLA) at the Max Planck Institute for Psycholinguistics provides a unique record of how people around the world use language in everyday life. It focuses on collecting spoken and signed language materials in audio and video form along with transcriptions, analyses, annotations and other types of relevant material (e.g. photos, accompanying notes). Besides these collections, the archive also maintains other collections of language-related research data from the various (current and former) departments and research groups within the Max Planck Institute for Psycholinguistics. This includes data resulting from neurobiological studies, genetics studies, and behavioural studies.

TLA's host organisation, the Max Planck Institute for Psycholinguistics, is part of the Max Planck Society legal entity (formally "Max-Planck-Gesellschaft zur Förderung der Wissenschaften e. V."), which is a registered association under German law. The Max Planck Society is funded by the German federal government as well as the individual German states where the 84 Max Planck Institutes are located.

The Language Archive:
https://archive.mpi.nl [accessed 05-11-2021]

Max Planck Institute for Psycholinguistics:
https://www.mpi.nl [accessed 05-11-2021]

Max Planck Society:
https://www.mpg.de/en [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
accept

Brief Description of the Repository's Designated Community.

TLA mainly targets scholars of language-related research disciplines. This includes linguists, anthropologists and psychologists, but also neurobiologists and geneticists who study language-related phenomena. The rich collections of recordings of languages and cultures in their natural context however are of interest to a much broader range of scholars, e.g. ethnomusicologists and ethnobiologists, as well as to the language communities themselves and the general public.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

A. Content distributed as deposited, B. Basic curation – e.g. brief checking; addition of basic metadata or documentation, C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

The level of curation that TLA performs varies per collection but is never data-level curation. Most often materials are either distributed as deposited or basic curation is performed. If the deposited materials include sufficiently rich metadata descriptions and contain only files in accepted formats, content is distributed as deposited after performing a series of automated checks. In other cases, the archive staff may add additional metadata or ask the depositor to provide more metadata and/or documentation. In principle, the archive requires the data to be deposited in accepted formats and provides guidance to depositors on how to convert files into those formats. In case depositors are not able to perform the conversions themselves due to the complexity or due to the volume of the deposit, the archive staff may offer to perform the conversions for them.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Insource/Outsource Partners. If applicable, please list them.

TLA replicates its data to two data centres of the Max Planck Society, the MPCDF in Garching and the GWDG in Göttingen. These data centres have a services catalogue and an agreement with the society about the services that they offer to all Max Planck Institutes. Both centres provide their services to the society at their best effort without strict guarantees e.g. regarding uptime. In case of poor performance however the society will take action such that changes in procedures or management are implemented.

TLA is one of the centres of the CLARIN European research infrastructure. As such, TLA also makes use of a number of services offered by CLARIN, such as the Service Provider Federation to connect to national identity federations and the Nagios monitoring service (in addition to in-house Nagios monitoring of servers and services).

Max Planck Computing and Data Facility (MPCDF) services:
https://www.mpcdf.mpg.de/services/supercomputing [accessed 05-11-2021]

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) services catalogue:
https://www.gwdg.de/web/guest/about-us/catalog/catalog-mpg [accessed 05-11-2021]

CLARIN services:
https://www.clarin.eu/content/services [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Summary of Significant Changes Since Last Application (if applicable).

Access to the collections in the archive is now handled via two "portals", one labelled "The Language Archive" and one labelled "MPI for Psycholinguistics Archive". The first one focuses on collections of language recordings, whereas the second one focuses more on collections of other data resulting from language-related studies carried out at the Max Planck Institute for Psycholinguistics. There is a partial overlap of collections that appear in both portals. This change is purely at the presentation level, all data still sit in a single underlying Fedora repository and are subject to the same procedures and workflows.

One copy of the data at the GWDG data centre in Göttingen now uses S3-compatible storage, which can be accessed directly from the Versity HSM storage software. Accessing this remote copy can be faster than accessing a local copy on tape for smaller file sizes due to the long seek times of tape-based storage.

An “academic use” access level was introduced in order to automatically grant access to certain resources to anyone with an academic affiliation (as determined by their federated login or assigned manually for local accounts). Access levels are now also displayed in colour-coded labels for each file, as well as in a “filter” that allows users to select only materials with a certain access level.

A step towards linking more datasets to scholarly publications was made by an agreement between DANS and other research data repositories in The Netherlands to include a "fundingReferenceNumber" for a given project in the metadata in a standardised form. The syntax for this reference was taken from the OpenAIRE metadata guidelines. By doing this, the NARCIS "gateway to scholarly information in the Netherlands" that harvests large amounts of scholarly metadata and that is maintained by DANS is now able to show aggregated outcomes including publications and datasets for a given research project. At the moment this was only piloted for a single project at our end (see link below) but the goal is to add these reference number to other projects with Dutch funding in the future.

A citation function was implemented for the repository, which displays an automatically extracted citation from the available metadata in a given collection, a part of collection or a file within a collection.

Work has started on a migration trajectory towards the next versions of the Fedora/Islandora repository solution. This migration needs to be completed by the end of November 2022 at the latest, when the Drupal version 7 software that forms an important part of the solution will become End of Life and will therefore no longer receive security updates from that point onwards. The current CoreTrustSeal assessment is still based on the “legacy” versions of Fedora and Islandora though.

Data Archiving and Networked Services (DANS):
https://dans.knaw.nl [accessed 05-11-2021]

OpenAIRE metadata guidelines for funder reference:
https://guidelines.openaire.eu/en/latest/data/field_contributor.html#nameidentifier-ma-o [accessed 05-11-2021]

Example of a project page on NARCIS that includes data hosted at TLA:
https://www.narcis.nl/research/RecordID/OND1366397/Language/en [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

ORGANIZATIONAL INFRASTRUCTURE

1. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The Language Archive at the Max Planck Institute for Psycholinguistics (TLA) holds one of the largest collections of language related research data worldwide. It strives to provide a unique record of how people around the world speak in everyday family life. It focuses on collecting spoken and signed language materials in audio and video form along with transcriptions, analyses, annotations and other types of relevant ancillaries (e.g. photos, accompanying notes). The archive includes speech data from everyday interactions in families and communities, focusing on families with children, but including naturalistic data from adult conversations from under-studied languages and linguistic phenomena. Due to a variety of reasons, including globalization and political repression, many of these under-studied languages are at risk of no longer being spoken by future generations. TLA's goal is to preserve its language collections for future use and to make them available for research and other uses both now and in the future. A selection of the archive's holdings was recognized by UNESCO as Memory of the World in 2015.

In addition to collections of language materials, the archive also holds a range of materials from language-related studies conducted by research staff at the Max Planck Institute. This includes e.g. eye tracking data, fMRI brain scans, gene sequences, and reaction-time measurements. For these types of data, there is a minimum preservation commitment for 10 years after publication, which is a requirement from the Max Planck Society. There are currently no policies to deaccession these materials after this period, but this standpoint will be re-evaluated from time to time.

A public mission statement which is endorsed by the directorate of the Max Planck Institute for Psycholinguistics can be found here:
https://archive.mpi.nl/mission [accessed 05-11-2021]

UNESCO Memory of the World entry:
https://en.unesco.org/memoryoftheworld/registry/258 [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

2. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

Depositors within the MPI for Psycholinguistics are required to deposit any research data that resulted in a publication in the archive. Since these data belong to the Max Planck Institute for Psycholinguistics, no specific deposit agreements are needed.

TLA uses deposit forms for external depositors. For depositors within the DOBES (DOcumentation BEdrohter Sprachen/ Documentation of Endangered Languages) programme, this is the following agreement: http://dobes.mpi.nl/ethical_legal_aspects/DOBES-daa-v1.pdf Agreements with other external depositors are largely based on this.

The archive uses a number of agreements for usage of its collections, including:
- CLARIN RES+BY+NORED
- CLARIN ACA+BY+NORED
- Creative Commons BY-NC-SA 3.0
- The DOBES Code of Conduct v.2, which covers all collections that are part of the DOBES part of the archive.

In addition, there are a number of collection-specific usage agreements that contain for example a specific citation requirement or a co-authorship requirement.

Part of a funded project that is mentioned in more detail under R3 will be the assessment and possible revision of the legal and ethical agreements (including licenses) that are currently in place, taking some developments that have taken place in recent years into account, including the General Data Protection Regulation (GDPR) of the European Commission and the CARE Principles for Indigenous Data Governance.

The archive has 4 access levels for granting/restricting access to archived resources. The depositor decides which level is appropriate for their collection, or part of their collection:
- Materials that are completely open and can be downloaded by anyone without the need to log in
- Materials that are available to any authenticated user. Users can log in with their own institutional account if their organization is part of the CLARIN identity federation or can create a dedicated account for the archive otherwise. These accounts for the archive are manually verified before they become active to ensure that they are from genuine people with an interest in the archive.
- Materials that are accessible to all users with an academic affiliation. This is either determined automatically in case users log in with their own institutional account, or manually for users who create an account specifically for the archive.
- Materials for which access needs to be requested. In many cases such a request will need to be approved by the

depositor or by a designated organization.

Depositor-Archivist agreement within the DOBES programme:
http://dobes.mpi.nl/ethical_legal_aspects/DOBES-daa-v1.pdf [accessed 05-11-2021]

DOBES Code of Conduct v2:
https://dobes.mpi.nl/ethical_legal_aspects/DOBES-coc-v2.pdf [accessed 05-11-2021]

CLARIN end user license agreements:
https://www.clarin.eu/content/licenses-and-clarin-categories [accessed 05-11-2021]

General Data Protection Regulation (GDPR):
https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en [accessed 05-11-2021]

CARE Principles for Indigenous Data Governance:
https://www.gida-global.org/care [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

3. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
Accept

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept

Response:

TLA aims to preserve and provide access to its language collections indefinitely, and for a minimal period of 10 years for its other collection types. The core of the archive has a solid basis in terms of funding and permanent staff within the MPI for Psycholinguistics as described under R5. The MPI as a whole gets reviewed every three years by its international scientific advisory board. Every 6 years this review also includes a review of the technical support group, to which the core of the archive staff belongs. In theory, Max Planck departments or even entire institutes could be closed down after a number of unfavourable reviews in a row. The MPI for Psycholinguistics has until now always received outstanding reviews though and can therefore be considered a solid organisation to host the archive.

Should the archive no longer be able to fulfil its task within the current host organization however, it will do its utmost to transfer the collections to a different organization that can take over the long-term preservation and access task, either within the Max Planck Society, or elsewhere such as with another CLARIN centre. The Max Planck Society guarantees bit-stream preservation of TLA’s holdings for a period of 50 years.

The MPI has secured funding from the Volkswagen Foundation together with the Institute for the German Language (IDS) in Mannheim to work on the establishment of an accessible copy of the DOBES endangered languages collections at the IDS. This two-year project will start in 2022 and will include work on the technical, organisational and legal aspects that play a role for the creation of this accessible copy. The IDS is part of the newly funded “Text+” research data infrastructure consortium in Germany, within which the IDS as well as other partners have the necessary expertise to take care of long-term preservation of the DOBES materials in case the MPI would no longer be able to fulfil this task. The outcomes of this project could be used for other parts of the archive as well.

Work on providing a fall-back instance of the repository system at the Max Planck Society’s data centres where the replication copies are located was postponed until after the migration to the newer Fedora and Islandora versions. The new setup will be deployed using Docker containers and a high degree of automation, which makes the establishment of further instances a lot easier.

Max Planck Institute for Psycholinguistics Advisory Boards:
https://www.mpi.nl/page/advisory-boards [accessed 05-11-2021]

Text+ consortium:
https://www.text-plus.org/en/home/ [accessed 05-11-2021]

Volkswagen Foundation:
https://www.volkswagenstiftung.de/en [accessed 05-11-2021]

Leibniz-Institut für Deutsche Sprache (IDS):

https://www.ids-mannheim.de [accessed 05-11-2021]

Docker:

https://www.docker.com [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

4. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

TLA contains various kinds of datasets, for which different legal and ethical criteria play a role. Researchers within the Max Planck Institute for Psycholinguistics who use human subjects in their studies are bound to the ethical rules regarding human subject data from the Max Planck Society and have to get approval from an ethics committee at Radboud

University or at Radboud University Medical Center, depending on the kind of study. A large collection within TLA is the DOBES (DOcumentation BEdrohter Sprachen) archive, which contains data sets that are collected within the DOBES endangered languages documentation programme funded by the Volkswagen Foundation. Depositors in a DOBES project are bound to ethical rules of the DOBES code of conduct. The archive does not (and cannot) systematically verify whether the data it receives is collected according to these rules.

Data with disclosure risk are typically only available upon request (see R2). Physical access to the archive's technical infrastructure is limited to only those people who manage the infrastructure (system administrators). Besides that, archive management and curation staff are the only ones who have access to all data. All staff of the Max Planck Society need to follow the data protection guidelines of the society (available upon request). All technical staff at the MPI sign a data confidentiality agreement upon the start of their employment.

Archived data are anonymised whenever appropriate and possible, however in the case of audio and video recordings, no transformations are performed to make subjects unrecognisable, as this would render the data useless for many purposes. In some cases, data are explicitly not anonymised if this is the wish of the subject.

The Max Planck Institute for Psycholinguistics now has a Data Protection Committee that is tasked with ensuring GDPR-compliance for all personal data that is collected and processed within the institute. This is done in close collaboration with the Data Protection Officer of the Max Planck Society head office.

Max Planck Institute for Psycholinguistics Data Protection Committee:
https://www.mpi.nl/page/data-protection-committee [accessed 05-11-2021]

Data Protection Officer of the Max Planck Society:
https://www.mpg.de/officers-of-the-mpg [accessed 05-11-2021]

Ethics committee for research involving human subjects at Radboud University Medical Center. (Dutch only, but instructions available in English as well):
https://www.radboudumc.nl/over-het-radboudumc/kwaliteit-en-veiligheid/commissie-mensgebonden-onderzoek [accessed 05-11-2021]

Informed consent and ethics committees at Radboud University:
https://www.ru.nl/rdm/collecting-data/informed-consent-ethics-committees [accessed 05-11-2021]

Description of the DOBES programme for documentation of endangered languages:
https://dobes.mpi.nl/dobesprogramme/ [accessed 05-11-2021]

DOBES Code of Conduct v2:
https://dobes.mpi.nl/ethical_legal_aspects/DOBES-coc-v2.pdf [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

5. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The core of TLA is funded by the Max Planck Institute for Psycholinguistics, with additional project funding typically coming from Dutch, German or European research funders (currently from the Dutch Research Council NWO and the Volkswagen Foundation). At the moment, the archive has the following staff:

- 1 Senior Archive Manager, 1 FTE, permanent position. Overall management, policy development, technical infrastructure design and planning, repository system management, data/metadata curation, user assistance and training
- 1 Junior Archive Manager, 1 FTE, permanent position. User assistance & training, digitisation, file/format conversions, data/metadata curation, user manuals & documentation
- 1 Junior Archive Manager, 1 FTE, initial 5-year position extended with one year (2017-2023). See above.
- 1 Software Developer back-end: 0.6 FTE, permanent position. Development and maintenance of back-end components in relation to the repository, mainly in Java.

- 1 Software Developer front-end: 1 FTE, 5-year position (2018-2023). Development and maintenance of front-end components in relation to the repository, mainly in PHP.
- 1 Software Developer applications: 0.8 FTE, permanent position. Development and maintenance of stand-alone client applications, mainly the ELAN linguistic annotator, written in Java.
- 1 Software Developer applications: 0.8 FTE, 2.5-year position (2021-2023). Modularisation and restructuring of the ELAN linguistic annotator codebase, interoperability with the CLARIAH infrastructure.

In addition to these positions, the archive makes use of the IT facilities and staff of the Max Planck Institute for its servers, storage, and general infrastructure support. The development of the customised parts of the archive’s repository system is done in collaboration with the Meertens institute in Amsterdam.

The management team that is responsible for the archive consists of the Director of the MPI’s Language Development department, the head of the MPI’s technical group, and the Senior Archive Manager.

The core operation and development of the archive is part of the overall technical infrastructure that the MPI's technical group maintains. This is a priority for the long term. Any additional project-based work to enhance the archive's exploration environment and to support additional external deposits currently falls under the responsibility of the director of the Language Development Department, who is committed to apply for additional funding when necessary.

Training and professional development wishes from the archives staff can almost always be accommodated by the MPI.

CLARIAH (Common Lab Research Infrastructure for the Arts and Humanities) Netherlands:
https://www.clariah.nl [accessed 05-11-2021]

The Language Archive team:
https://archive.mpi.nl/tla/people [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

6. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific

guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The senior archive manager closely follows developments in the areas of research data, data preservation and repository systems, by attending conferences and workshops, by monitoring online sources of information, and by being a member of the Research Data Alliance.

The software developers are following relevant developments in their areas of expertise by attending workshops, development camps or conferences, in addition to monitoring relevant online sources of information.

As a CLARIN Centre, the archive is informed about technical developments in the field of language resources, tools and infrastructure, and is periodically evaluated by the CLARIN centre committee. The CTS assessment is an important source of information for this periodic evaluation.

TLA is part of the DELAMAN network of endangered languages archives, which is platform for exchanging knowledge between archives within this domain and for providing information to people involved in the documentation of endangered languages.

The directorate of the MPI as well as the MPI’s Research Facilities Committee (RFC) inform the archive about scientific requirements form the different sub-disciplines that are present in the institute. The RFC meet four times per year and inform the archive staff in case they see any need for changes regarding workflows or functionality of the repository.

Direct feedback from users and depositors is obtained via a “feedback button” on the bottom right of the archive’s portals, via the contact forms and via email.

Technical advice about IT infrastructure is available from the Max Planck Society’s head office. They also conduct periodic

IT audits of all member institutes.

Research Data Alliance (RDA):

https://www.rd-alliance.org [accessed 05-11-2021]

DELAMAN network:

https://www.delaman.org [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accepted response

DIGITAL OBJECT MANAGEMENT

7. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

TLA keeps MD5 checksums for every deposited object in its Fedora Commons (3.8.1) repository. In addition, it uses a Versity (SAM-QFS based) hierarchical storage management system that also maintains MD5 checksums in its file system inode metadata. The Fedora checksums are automatically validated periodically, and the file system checksums are compared upon each file access.

The archive allows changes to any metadata or data object. An ingest of a modified object will trigger the creation of a new version in the repository. All older versions are kept, and the Handle persistent identifiers will keep referring to them. The new version of the object will get a new Handle PID, as will any “ancestors” of the object in a hierarchical collection. This will allow users to refer to a “snapshot” of the collection at a given point in time.

An audit trail of all modifications is kept in Fedora Commons’ metadata.

Almost all data enters the repository via the self-deposit web interface. Users need to log in with their personal account and are being assigned deposit permissions for a specific section of the repository by the archive management staff.

Versity Software, Inc.:
http://www.versity.com [accessed 05-11-2021]

Fedora Commons 3.8.1 documentation:
https://wiki.lyrasis.org/display/FEDORA38/Fedora+3.8+Documentation [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

8. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

TLA's collection development policy can be found here: https://archive.mpi.nl/collection-development-policy. TLA has a list of accepted long-term archival formats as well as a list of accepted formats for which only bit-stream preservation can be guaranteed (https://archive.mpi.nl/accepted-file-formats). The latter list exists because for certain data types there are no suitable formats that fulfil the preservation format criteria of TLA's preservation policy. Yet, the archive has the obligation to preserve all data that is collected within the MPI and that resulted in a publication for a
minimum duration of 10 years, as determined by the Max Planck Society. Formats that are not on the list will be rejected and should be converted first, in principal by the depositor (the archive's staff may assist in exceptional cases). The archive performs automatic file format validations upon ingest. Unaccepted formats as well as invalid files in accepted formats are rejected. The archive also enforces file extension conventions as listed in the accepted formats list, files that do not conform to these conventions are rejected.

For certain metadata profiles, a number of metadata fields are mandatory. In other cases, archive staff will randomly verify whether the deposited metadata is acceptable. What is acceptable also depends on the type of collection, e.g. in the case of legacy data collections it would be more acceptable if the metadata is not complete than for recent materials of which the collector is still alive. For legacy materials, the archive's staff will try to enrich the received metadata with any further information that can be reliably recuperated.

Some examples of good metadata descriptions with an adequate level of detail:
(click "expand all" under "detailed metadata" to see all fields)
https://hdl.handle.net/1839/00-0000-0000-0016-7AB3-1
https://hdl.handle.net/1839/00-0000-0000-0008-3CD7-0

TLA collection development policy:
https://archive.mpi.nl/collection-development-policy [accessed 05-11-2021]

TLA accepted file formats:
https://archive.mpi.nl/accepted-file-formats [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:

Accept, however it would be nice to include links to one or two good examples of metadata for and archived object.
Provided thank you

9. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The archive uses a Versity (SAM-QFS based) hierarchical storage management system for its archival storage. This system consists of layers of different storage technology (SSD, Hard drive arrays, LTO7 tape library). A rule-based system determines which data is stored on which layer, when data moves from one layer to another, and how many copies of each file are stored. For all materials in the archive, one copy is kept on disk and two copies are kept on tape automatically.

All archived files are replicated to two data centres of the Max Planck Society, the MPCDF in Garching and the GWDG in Göttingen, which means that a minimum of 5 copies exist in 3 geographically distinct locations. All replicated copies elsewhere are stored under the agreement that they are made available under the same conditions and access restrictions, if they are made available (which is not yet the case). One copy at the GWDG data centre is stored on a storage system that is compatible with the Amazon S3 protocol. This copy can be accessed directly from the Versity storage system software and therefore serves as a 4th copy within this setup. This “S3 copy” can often be accessed quicker than the local copies residing on tape, depending on the size of the file.

The MPI for Psycholinguistics has a disaster recovery plan as part of its general IT operations handbook. However, in the case of a major disaster such as fire or flooding, it may take a long time before the physical structures are back in place to execute the plan. Therefore, the archive will be working on replicating its repository software to the data centres as well, such that access can be provided via these locations almost instantly. Work on this will begin once the migration to the newer repository software version is completed, which is expected to be in the fall of 2022, and should be completed in the course of 2023.

Versity Software, Inc.:
http://www.versity.com [accessed 05-11-2021]

Max Planck Computing and Data Facility (MPCDF):
http://www.mpcdf.mpg.de [accessed 05-11-2021]

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG):
https://www.gwdg.de [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

10. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

TLA’s preservation policy can be found here: https://archive.mpi.nl/preservation-policy

As mentioned under R1 and R3, TLA has different preservation commitments for different types of collections. For the largest part of its holdings, the collections with spoken, signed, or written language data, the goal is to preserve these indefinitely. For the other types of collections, the minimum preservation commitment is 10 years after publication. There are currently no policies to deaccession these materials after this period, but this standpoint will be re-evaluated from time to time.

The archive’s rights to archive, transform and provide access to its holdings are arranged in the deposit agreements as far as external depositors is concerned.

Any future file format conversions will be done with the most appropriate tool for the job at that time, ideally without loss of information. An important criterium for choosing a tool is the possibility to run it in a batch-like fashion, which typically means that command line tools are preferred. The ffmpeg video encoding tool e.g. would be our tool of choice at the moment for any video file format conversions. Future file format migrations will most likely be done without the involvement from individual depositors. As with current file format conversions that are done as part of the curation task, manual random quality checks will be performed for any future file format conversions as well.

TLA preservation policy:
https://archive.mpi.nl/preservation-policy [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

11. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

Technical validity is verified automatically for all deposited data and metadata upon ingest. The FITS File Information Tool Set which is produced by Harvard Library and which includes an array of tools and libraries to identify and validate various file formats is used to generate a detailed report for each deposited file. This report is automatically processed with a set of rules for each accepted file format to determine whether or not a file is accepted. Invalid files or files that do not fulfil all defined criteria are automatically rejected. The technical quality of the deposited data and metadata is randomly checked by the archive's staff after ingest. Archive staff deploy scripts to periodically check for obvious errors in metadata values across the archive, such as typos or empty values for important fields.

Many collections in TLA are created over long periods of time, i.e. depositors keep adding additional materials or newer versions of already archived materials. As such, it's not really possible to assess whether such collections are "complete". The usability of a multi-media language corpus for other researchers often depends on whether or not a reasonable portion of it has been transcribed, translated and further linguistically annotated. This is an incredibly time-consuming task though, so it often takes years after an initial deposit before these additional resources become available.

In some cases, publications that resulted from the deposited data are part of the collection itself. In other cases, links to publications are provided and methodology descriptions are included with the data. There are however still many collections for which publications exist, but no links are provided yet. As mentioned in R0, a step towards linking more data to publications was implemented by including a funding reference number in a standardised manner, such that the NARCIS "gateway to scholarly information in the Netherlands" is able to show all outcomes for a given project.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

12. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

Almost all deposits into TLA are done by researchers themselves via a web-based deposit interface. This interface enforces a rather strict workflow in terms of the structure of the deposit, the assurance that each file is linked to a metadata record, the assurance that each file is valid and is in an accepted format, and the specification of access permissions. TLA provides deposit manuals for internal and external depositors.

The deposit interface does offer some degrees of flexibility. The depositor can decide how to structure their deposit in terms of sub-collections within their collection, and they can choose whether to provide their metadata using web forms or as uploaded files. For users within the Max Planck Institute, the deposit interface can retrieve files from a user’s personal network share. For external users, files are uploaded and made available to the deposit tool via a Nextcloud instance hosted by the archive.

After passing a number of automated checks, the data is finally ingested into the Fedora repository. At that point, derivative files for dissemination are automatically generated for certain file types (currently for video, audio and images).

Ingests done by the archive's staff are typically also done via the same web-based deposit interface. Only in exceptional cases, command line batch ingest scripts are used. These scripts also make use of the same automated checks that the web-based deposit interface deploys.

Nextcloud:
https://nextcloud.com [accessed 05-11-2021]

Deposit manuals:

https://archive.mpi.nl/tla/deposit-manual [accessed 05-11-2021]

https://archive.mpi.nl/mpi/deposit-manual [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

13. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

TLA offers an elaborate search interface for the metadata of its holdings (https://archive.mpi.nl, choose “browser archive”) There is a free text search box as well as a faceted search option for the most important metadata fields. With this faceted search, users can either include or exclude values (exclude by using the minus sign behind the value). In addition to the facets, there are also alphabetically ordered lists of the facet values that can be accessed via the “Browse by” menu at the top.

TLA makes its metadata records available for harvesting via OAI-PMH in three different formats: the original CMDI metadata, a transformation to OAI-DC, and a transformation to OLAC (http://www.language-archives.org). The metadata are currently being harvested by the CLARIN Virtual Language Observatory (https://vlo.clarin.eu), the OLAC metadata aggregator (http://search.language-archives.org) and NARCIS (https://narcis.nl).

TLA makes use of the Handle system for its persistent identifiers. It runs its own Handle server instance for this purpose. Handle PIDs are issued for each individual object (data files, metadata files, and collection nodes). New versions of objects get a new PID, the old PID stays with that old version. When a child object of a collection or sub-collection is modified, the parent objects also get a new PID, up to the top-node of the collection.

A data citation function that generates a citation at the click of a button for each item that has a PID was implemented. The resulting citation is not always complete or accurate though, which gives us an indication of where additional metadata curation is needed (e.g. different variants of authors' names, missing dates, etc.).

Certain data types can be viewed directly in the browser in case the user has permissions to access the files. Online viewers are currently available for audio files, video files, images, annotated audio/video files, PDF files, Plain Text files, HTML files, and 3D objects.

Access interface to the archive including search:
https://archive.mpi.nl [accessed 05-11-2021]

OLAC metadata standard:
http://www.language-archives.org [accessed 05-11-2021]

CLARIN Virtual Language Observatory:
https://vlo.clarin.eu [accessed 05-11-2021]

OLAC metadata aggregator:
http://search.language-archives.org [site error when accessed on 05-11-2021]

Handle PID system:
https://www.handle.net [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

14. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

TLA requires metadata to be provided in one of several accepted CMDI profiles. CMDI (Component MetaData Infrastructure) is a metadata framework that is developed by the CLARIN European research infrastructure (https://www.clarin.eu, https://www.clarin.eu/content/component-metadata). It is not one fixed metadata schema, but instead a framework that makes it possible to compose metadata profiles from re-usable metadata component blocks. TLA only accepts a limited number of these profiles, as listed on this page: https://archive.mpi.nl/accepted-metadata.

The accepted data formats of the archive a chosen such that they are ideally both suitable for preservation, as well as usable today by the research community. For some formats, derivatives with a smaller size are generated for easier download and use (currently for video, audio and images).

TLA monitors the suitability of its accepted formats both in terms of long-term preservation as well as current-day use within the community. As mentioned in the preservation policy (https://archive.mpi.nl/preservation-policy), important sources of information that are consulted are the Library of Congress Sustainability of Digital Formats site and the National Archives current format summary. Due to the limited number of accepted archival formats that were strategically chosen, only one format migration has been necessary during the almost 20 years that the archive is in operation. TLA has the expertise in-house to develop automated migration pipelines, should the need for further file format migrations arise. In 2016, all metadata were migrated from the IMDI standard that was used up to that point to CMDI metadata.

TLA strongly advises depositors to include all necessary information that is needed to interpret the dataset with their deposit. This includes e.g. specification about annotation conventions, methodology reports, experimental setup reports, descriptions of equipment used, etc.

CLARIN ERIC website:
https://www.clarin.eu [accessed 05-11-2021]

CMDI component metadata:
https://www.clarin.eu/content/component-metadata [accessed 05-11-2021]

TLA accepted CMDI metadata profiles:
https://archive.mpi.nl/accepted-metadata [accessed 05-11-2021]

TLA preservation policy:
https://archive.mpi.nl/preservation-policy [accessed 05-11-2021]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

TECHNOLOGY

15. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

TLA uses Fedora Commons (3.8.1) for the core of its repository, on top of which it uses the Islandora front-end and an in-house developed deposit front- and back-end. It has also developed a number of additional modules for visualising certain data types (e.g. transcribed media, 3D models) and for extending Islandora's functionality (e.g. a citation generation module). Together, these components implement all technical functions of the OAIS reference model. TLA does not implement “Packages” in a literal sense, i.e. objects are stored as individual files with relation information being stored in Fedora Common’s metadata as well as in the CMDI metadata.

The development of some of the additional components that are used alongside Fedora/Islandora, such as the deposit back-end ("Doorkeeper" workflow engine) and a module to display the CMDI metadata is done together with the Meertens institute. All code is publicly available on GitHub.

As mentioned in R0, work is currently ongoing to migrate the repository to the latest versions of the Fedora and Islandora repository solutions. This migration will need to be completed by the end of November 2022 at the latest.

In terms of operating systems and hardware, the archive’s application servers are VMWare virtual machines running on the VMWare ESXi platform, which is running on Hewlett Packard hardware. The servers use SuSE Linux Enterprise Server as their operating system. The Versity storage system that is used for archival storage runs on Hewlett Packard servers with a CentOS operating system, which connect to a variety of storage hardware.

Fedora Commons 3.8.1 documentation:
https://wiki.lyrasis.org/display/FEDORA38/Fedora+3.8+Documentation [accessed 05-11-2021]

Fedora Commons 6.x documentation:
https://wiki.lyrasis.org/display/FEDORA6x/Fedora+6.x+Documentation [accessed 05-11-2021]

Islandora:
https://islandora.ca [accessed 05-11-2021]

Islandora latest version:
https://github.com/Islandora/islandora [accessed 05-11-2021]

OAIS reference model:

https://public.ccsds.org/pubs/650x0m2.pdf [accessed 05-11-2021]

TLA on Github:

https://github.com/TheLanguageArchive [accessed 09-02-2022]

https://github.com/TLA-FLAT [accessed 09-02-2022]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

16. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The Max Planck Institute for Psycholinguistics has an IT operations handbook that also described disaster recovery procedures and responsibilities. This handbook is in line with the ISO 27001 information security standard and is audited by the head office of the Max Planck Society.

Operating systems and software are kept as up-to-date as possible. A next generation firewall blocks access to most servers and ports for users outside the local network and for unregistered computers on the local network. Only services that need to be accessed by outside users (such as the web front-end of the archive) are passed through the firewall.

SURF, the organisation that provides the network infrastructure for educational and research institutions in The Netherlands, actively monitors its connections and reports to the institutions in case that vulnerable services, infected machines or attacks from outside are detected.

Most systems are implemented in a redundant manner such that single failures will not impact the operation of the infrastructure. Should any service fail nonetheless, the IT operations handbook describes which procedures need to be followed by whom in order to get the service back in operation as quickly as possible. The MPI employs two monitoring systems (Munin and Nagios) to monitor IT hardware and services. These systems inform the system management staff via SMS messages in case of critical issues. In addition, some of the archive's services are monitored via a Nagios instance under control of CLARIN. This Nagios instance runs at a computing centre in Juelich, Germany, and can therefore also detect connectivity issues from outside the MPI's network, which may go unnoticed by the internal monitoring instances.

As mentioned under R3, TLA will be working towards setting up additional instances of its repository software at the data centres that have the replication copies of the data. This will provide an almost instantaneous failover solution for providing access to the repository in case of a major disaster.

The Max Planck Institute for Psycholinguistics as a whole has undergone a risk assessment in terms of safety and security at the end of 2018. Apart from a few minor recommendations that were followed up, the assessment did not reveal any shortcomings. It is not known at this point when the next audit will be.

ISO 27001 requirements for an information security management system:
https://www.iso.org/isoiec-27001-information-security.html [accessed 05-11-2021]

Nagios IT infrastructure monitoring system:
https://www.nagios.com [accessed 05-11-2021]

Munin resource monitoring system:
https://munin-monitoring.org [accessed 05-11-2021]

SURF "SURFcert" security incidence support:
https://wiki.surfnet.nl/display/SURFcert/Dienstbeschrijving+SURFcert [accessed 09-02-2022]

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Response Acceptable

APPLICANT FEEDBACK

Comments/feedback

These Requirements are not seen as final, and we value your input to improve the CoreTrustSeal certification procedure. Any comments on the quality of the Requirements, their relevance to your organization, or any other contribution, will be considered as part of future iterations.

Response:

Reviewer Entry

Reviewer 1

Comments:
This is an excellent, short and concise application for CTS renewal. Very few changes from the original 2019 file, some technical details have been added which could be useful for other repositories looking for examples of a successful implementation. R3 (continuity of access) is the only requirement which remains under the maximum rating, a path for improvement will be scrutinized for the next renewal.

Reviewer 2

Comments:
A very good well written document.