LDC Catalog

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:

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, Publication repository, Research project repository

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept.

Brief Description of Repository

The Linguistic Data Consortium (LDC) is an open consortium of universities, libraries, corporations and government research laboratories hosted by the University of Pennsylvania, Philadelphia, Pennsylvania USA. It was formed in 1992 to address the critical data shortage then facing language technology research and development. The US Advanced Research Projects Agency provided seed funding for the Consortium and the US National Science Foundation provided additional support via Grant IRI-9528587 from the Information and Intelligent Systems division. The National Institute of Standards and Technology also provided early support.
Initially, LDC's primary role was as a repository and distribution point for language resources. Since that time, and with the help of its members, LDC has grown into an organization that creates and distributes a wide array of language resources. LDC also supports sponsored research programs and language-based technology evaluations by providing resources and contributing organizational expertise with support from, and in collaboration with, a wide range of international organizations in the commercial, non-profit sectors and government sectors, including sponsors within the US Departments of Commerce, Defense, Education, Homeland Security, Interior, Justice and Treasury.
LDC has distributed more than 200,000 copies of over 3,000 databases covering 90 different languages to more than 7,000 organizations in over 100 countries.

URL: https://www.ldc.upenn.edu/about (About LDC) (04/02/2021)

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept.

Brief Description of the Repository's Designated Community.

LDC's community includes linguists, computer scientists, social scientists and others engaged in language-related research, education and technology development.

URL: https://www.ldc.upenn.edu/about (About LDC) (04/02/2021)

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, D. Data-level curation – as in C above; but with additional editing of deposited data for accuracy

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept.

Comments

The level of curation provided by LDC depends on the particular circumstance, ranging from use as a technology evaluation to publication in the public catalog. As mentioned above, LDC supports common task evaluation organized by the community. Data sets are distributed to a restricted set of task participants. This is typically in the form of content as deposited with LDC since that data has already been prepared by task organizers for the specific task or evaluation. Data must remain in this format to serve as a benchmark and support subsequent comparison of new approaches and algorithms on known stable data. In this case, task organizers provide data to LDC and we, in turn, provide distribution and archival services without alteration. Distribution is limited to only those participating in the evaluation. In the event the data is released into our public catalog at a later date, it receives a higher level of curation as described below.
For data released into LDC's public catalog, LDC applies basic, enhanced and data level curation depending on the condition of the corpus deposited. Once a publication proposal has been accepted, LDC staff works with providers to facilitate delivery of the data and performs extensive quality assurance to ensure that data is complete, error free and ready to use. Those activities include data format checks, directory structure and documentation reviews and applying descriptive metadata schema. LDC manages and monitors archived data applying updates and bug fixes as needed and coordinating migration to new formats and platforms.

URL: https://www.ldc.upenn.edu/data-management/providing-data (Providing Data & Associated Subpages) (04/02/2021)
URL: https://www.ldc.upenn.edu/data-management/curation-distribution (Curation and Distribution Services (04/02/2021)

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept.

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

LDC delivers some data from Amazon's AWS cloud services. LDC uploads data to AWS and provides unique signed URLs for encrypted download to users. This is a closed loop between LDC and the cloud (part of the University of Pennsylvania's Amazon tier) and complies with the University's data security protocols. For disaster recovery purposes, LDC also backs up data to Amazon in addition to its backup on local storage.

URL: https://www.isc.upenn.edu/infosecpolicy20100308-computersecurity (Data Security Policy) (04/02/2021)

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept.

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

Since LDC's certification in 2018, we have enacted the following signification changes:
• Through membership with DataCite, we have assigned Digital Object Identifiers (DOI) to all our public data sets and continue to assign them to new releases.
• We enhanced our catalog's metadata and corpus discoverability by adding "Related Works", an interlinking field to connected related resources based on a set of controlled vocabulary. The Related Works schema is based on a taxonomy developed by META-SHARE and LDC with modifications relevant to LDC's language resources.
• We developed and implemented more specific technical and documentation guidelines with new requirements to increase the initial quality of submitted data
• Set guidelines to transition uncirculated data to a lower, but equally secure, storage tier to more efficiently manage our data holdings.
• Reconfigured digital delivery infrastructure to optimize datasets and increase efficiency of storage.

• Improved data submission with LDC Submissions, an in-house platform that provides infrastructure and resources for describing and uploading data sets and communicating between data providers and LDC.

URLs
https://www.ldc.upenn.edu/news/ldc-adds-dois-catalog (LDC adds DOIs to the Catalog) (04/02/2021)
https://www.ldc.upenn.edu/sites/default/files/ldc-related-works-schema.pdf (04/02/2021)
https://www.ldc.upenn.edu/data-management/using (Using LDC Data) (04/02/2021)
https://www.ldc.upenn.edu/data-management/providing-data (Providing Data) (04/02/2021)
https://www.ldc.upenn.edu/data-management/providing/publication-process (Publication Process) (04/02/2021)
https://www.ldc.upenn.edu/data-management/providing-data/technical-guidelines (Technical Guidelines) (04/02/2021)
https://www.ldc.upenn.edu/data-management/providing/documentation-guidelines (Documentation Guidelines) (04/02/2021)
https://www.ldc.upenn.edu/about/facilities/it-networking (IT Infrastructure) (04/02/2021)
https://submissions.ldc.upenn.edu/ (LDC Submissions) (04/02/2021)

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 4

Response:

Preservation is implied in all aspects of LDC's operations executing its mission and is the ultimate objective of all processes relating to data submission, review, publication, curation, and archiving.
The impetus for LDC's founding in the early 1990s was the need for a permanent archive to distribute large volumes of data that could be shared and re-used in support of human language technology development. This goal could only be accomplished if LDC data was identified, archived and stored so as to preserve its original form as published. This is reflected in the Consortium's distribution and archiving practices from 1993 to the present.
From the first data deposits that appeared with the Catalog's launch in 1993 to the latest publication, LDC data is characterized by:
(1) persistent identifiers, transparent descriptions and descriptive metadata; data formats as they were originally generated (normalized and updated as needed for perseverance), and straightforward user licenses, all to support broad user accessibility
(2) public guidelines for preparing and submitting data for publication, internal workflows for data intake, review, curation and long-term archiving, all to support data preservation

URL: https://www.ldc.upenn.edu/about/mission (Mission) (04/02/2021)
https://www.ldc.upenn.edu/about/ldc-overview (LDC Overview) (07/22/2021)
https://www.ldc.upenn.edu/data-management/providing-data (Providing Data) (07/22/2021)
https://www.ldc.upenn.edu/data-management/providing/publication-process (Publication Process) (07/22/2021)
https://submissions.ldc.upenn.edu/ (LDC Submissions) (07/22/2021)
https://catalog.ldc.upenn.edu/ (LDC Catalog) (07/22/2021)
https://www.ldc.upenn.edu/data-management/using/licensing (Licensing) (07/22/2021)
https://www.ldc.upenn.edu/data-management/using-data/user-agreements (User Agreements) (07/22/2021)
https://www.ldc.upenn.edu/data-management/preserving (Preserving Data) (11/24/2021)

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 4

Response:

LDC data may be used for language-related research, education and technology development. Data cannot be redistributed to others outside of the member/licensee organization/research group. These restrictions on data use are handled as follows:

• All LDC members and licensees sign agreements limiting and acknowledging the limitations on data use before they receive data.

• LDC's Membership Agreements regulate the use of data published in LDC's catalog by for-profit, not-for-profit and government entity members.

• Most public data licensed by non-members is governed by the LDC User Agreement for Non-Members.

• Certain LDC data sets are governed by corpus-specific license agreements which supersede the LDC Membership Agreements and the LDC User Agreement for Non-Members and must therefore be signed by all licensees (members and non-members).

• Data used in common task evaluations is usually governed by an evaluation license that limits the use of the data for the purposes of, and the duration of, the task.

• LDC handles legal, regulatory and contractual issues -- including intellectual property, human subjects protections and export controls -- during the publications submissions review process. It enters into distribution agreements with data providers that permit LDC to store and distribute the data under its typical model (subject to the membership agreement or non-member agreement) or under explicit use restrictions that will be communicated to users in a corpus-specific license.
• LDC's External Relation groups includes a contract expert who remains in regular contact with the University of Pennsylvania's (LDC's host organization) Office of General Counsel and Institutional Review Board for the treatment of human subjects and sits on a University advisory board for export controls.
• LDC maintains copies of all executed membership and license agreements.
Action in cases of noncompliance with user licensing requirements is determined on a case-by-case basis. Measures can range from corrective steps (e.g., requiring a user to retrieve copies of data distributed in violation of a license agreement) to removing a resource from a user's data queue to barring a user from licensing LDC data altogether.

URL: https://www.ldc.upenn.edu/members/agreements (Membership Agreements) (04/02/2021)
URL: https://catalog.ldc.upenn.edu/license/ldc-non-members-agreement.pdf (LDC User Agreement for Non-Members) (04/02/2021)
URL: https://www.ldc.upenn.edu/data-management/curation-distribution (Curation and Distribution Services) (04/02/2021)
URL: https://www.ldc.upenn.edu/data-management/using-data/user-agreements (User Agreements) (04/02/2021)
URL: https://www.ldc.upenn.edu/data-management/providing/publication-process (Publication Process) (04/02/2021)

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 3.

Response:

Copies of most LDC publications are available from the University of Pennsylvania's Van Pelt Library. Van Pelt creates catalog records for each publication deposited with corresponding descriptive, structural, and other appropriate metadata. The deposit procedure had been in abeyance pending changes to the Library's digital holdings framework. However, the Library has recently completed a reorganization process. We have resumed deposits of physical copies of LDC corpora for cataloguing and preservation with the release of September 2021. We will also backfill publications not deposited.
In the event that LDC terminates its operations, all Consortium property remains the property of the University of Pennsylvania. The University of Pennsylvania's Van Pelt Library maintains and provides access to LDC data as it does with respect to all of its holdings and will continue to do so in the event LDC is no longer in operation. Mechanisms for continued accessibility to users outside the University community are being considered for implementation . LDC Catalog metadata is mirrored in the Open Language Archives Community and the Government Catalog of Language Resources (for US government use).

URL: https://catalog.ldc.upenn.edu/ (LDC Catalog) (04/02/2021)
URL: http://www.library.upenn.edu/ (Franklin Catalog) (04/02/2021)

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 4

Response:

During the submissions review process, corpus providers describe how they comply with any legal or ethical regulations governing a particular data collection. Data providers sign an agreement asserting that they own and/or have rights to grant LDC permission to distribute. Data received by LDC for publication is reviewed by LDC staff for compliance with any applicable laws and regulations relating to data protection. LDC sometimes refuses to publish corpora if the creator cannot provide evidence that they have complied with applicable laws and regulations governing copyright, informed consent and the ethical treatment of human subjects.
Any special conditions, such as those related to data with disclosure risk, are handled in the user license developed for the corpus.
Personal identifying information is typically removed by data producers/corpus providers from any data collected from humans (unless participants otherwise consented to the release of that data). LDC staff is trained to handle the management of data with disclosure risk.
Ethical requirements for human subjects collections and the requirement that data providers must be able to provide LDC the right to store and distribute the submitted resource in a signed distribution agreement are described on the Publication Process page, https://www.ldc.upenn.edu/data-management/providing/publication-process. Submitters are also asked to provide relevant information regarding human subjects collections and data ownership/licensing constraints when they create a submission in LDC Submissions, https://submissions.ldc.upenn.edu/.
GDPR has not been an issue to date with respect to data sets reviewed for publication. As indicated above, researchers submitting data collected from human subjects must provide evidence of some approval for the collection and consent from subjects.
We review all submissions with reference to any laws or regulations for intellectual property, human subjects research and export controls. Some submissions may implicate other laws or regulatory schemes which we then review. There is no internal manual for this process.

URL: https://www.ldc.upenn.edu/data-management/providing-data/technical-guidelines (Technical Guidelines) (04/02/2021)
URL: https://www.ldc.upenn.edu/data-management/providing/documentation-guidelines (Documentation Guidelines) (04/02/2021)

URL: https://www.ldc.upenn.edu/data-management/curation-distribution (Curation and Distribution Services) (04/02/2021)

URL: https://submissions.ldc.upenn.edu/ (LDC Submissions) (04/02/2021)

URL: https://www.ldc.upenn.edu/data-management/providing/publication-process. (Publication Process) (07/22/2021)

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 4

Response:

LDC is hosted by the University of Pennsylvania and is a center within the University's School of Arts and Sciences. Funding for LDC's activities relating to data publication and distribution is supported by Consortium member fees and non-member data license fees.

LDC has sufficient funding and staff resources to operate in the long-term. We cannot share specific budget information.

LDC employs approximately 40 full-time staff. The External Relations group is composed of four full-time staff members whose responsibilities include membership, publications, data licensing and data delivery. The Consortium is housed in an office space of approximately 11,000 square feet. LDC maintains up-to date equipment for publication reparation and distribution. LDC staff are qualified to perform the duties of their positions and have access to ongoing training and professional development.

We can provide an organization chart only with an agreement that the chart cannot be part of our public submission.

URLs: https://www.ldc.upenn.edu/about (About) (04/02/2021)
https://www.ldc.upenn.edu/about/staff (Staff) (04/02/2021)
https://www.ldc.upenn.edu/collaborations/other (Other Collaborations) (04/02/2021)
https://www.ldc.upenn.edu/about/facilities (LDC Facilities) (04/02/2021)
https://www.ldc.upenn.edu/about/facilities/publications (Publications) (04/02/2021)

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 4

Response:

LDC consults with the University of Pennsylvania's IT departments on policies and issues affecting data storage, security and accessibility. LDC regularly seeks the feedback of its user community about Consortium data and services through surveys. The last user survey was conducted in 2020; there is no public document associated with the 2020 user survey. LDC follows developments in the digital resources community on issues affecting data distribution, curation and archiving including best practices for data storage and delivery. This includes attending conferences and workshops highlighting digital repository issues, such as those hosted by the Research Data Alliance and the European CLARIN program. As the first and most active language resource repository, LDC has been a leader in solutions for developing and maintaining digital archive collections and has developed a network of community connections with like-minded organizations. When needed, those long-time community connections can be accessed to seek additional guidance and feedback.
LDC is an active member of the global language resource community. For instance, LDC works with the European Language Resources Association, the Linguistic Data Consortium for Indian Languages, the South African Centre for Digital Language Resources, Gengo-Shigo-Kyokai and others regarding the role of data centers in language resource development and distribution.
LDC also collaborates with global networks including the British National Corpus Consortium, E-MELD, European projects such as CLARIN, ENABLER, FLaReNet and META-NET, the Japan-based Language Grid and the US TalkBank project.
LDC is a member of the Open Languages Archives Community (OLAC), an international partnership to create a worldwide virtual library of language resource metadata, which includes consensus for best practices for digital archiving. LDC's Catalog (searchable through OLAC) consistently receives OLAC's five-star rating for overall metadata quality.

URL: https://www.ldc.upenn.edu/data-management/data-center-distribution (Advantages of Data Center Distribution) (04/02/2021)

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

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 4

Response:

LDC generates checksums upon receipt of data and immediately places that data into its storage workflow which provides appropriate backup, replication and disaster recovery. Volumes have ongoing, built-in integrity checks to ensure fixity. Established update procedures ensure no data is lost and provide the ability to roll back to previous versions. Major updates to a corpus are given a new version number, indicated in the title of the corpus, and released as a new data set. Previous versions remain accessible under their existing catalog ID and URL. Minor updates are made in place and noted in an internal change log and external description; removed or altered data is archived along with checksums to ensure that any previous version can be re-created if need be. Physical backups of previous versions are also maintained. Providers are required to deliver checksums alongside their data. Permissions for data access are restricted to only the staff directly responsible for archival management. Providers' identifies are verified through the agreement process & discussion, involvement from the provider's institution, and document and data review.
URL: https://www.ldc.upenn.edu/about/facilities/it-networking (IT Infrastructure) (04/02/2021)
URL: https://www.ldc.upenn.edu/data-management/curation-distribution (Curation and
Distribution Services) (04/02/2021)
URL: https://www.ldc.upenn.edu/data-management/providing-data/technical-guidelines (Technical Guidelines) (04/02/2021)
URL: https://www.ldc.upenn.edu/data-management/providing-data/ documentation-guidelines (Documentation Guidelines) (04/02/2021)
URL: https://www.ldc.upenn.edu/data-management/using
(Using Data) (04/02/2021)

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 4

Response:

LDC provides public guidelines for data submission that include preferred filenames, standard formats and metadata preparation and instructions for preparing and submitting data for publication. Providers are required to submit metadata and other basic information about their data through LDC's website to trigger the submissions workflow. Further discussions about particular corpora continue with providers throughout the submissions and review process. If data is not in a preferred format, we talk to providers about the feasibility of conversion. If that is not possible, we ask the provider to submit documentation sufficient so that users across platforms and operating systems can use the data set for its intended purpose. When a submission is not deemed a good fit for LDC for either data or distribution issues, the provider is informed it has been declined and any data received is deleted from our servers. Very rarely are resources removed from the collection. In the small number of cases where this does happen, the data and catalog record are still stored internally, and the public facing entry is hidden or presented as a tombstone page based on DataCite's guidelines for use with DOIs. LDC's catalog entries provide descriptive metadata, including relations between works, following Dublin Core and OLAC standards. Resource documentation, including the catalog description, provides structural metadata (how to process/use the data set.) License information provides administrative metadata covering intellectual property, conditions on use and so on.
URLs https://submissions.ldc.upenn.edu/ (LDC Submissions) (04/02/2021)
https://www.ldc.upenn.edu/data-management/providing-data/technical-guidelines (Technical Guidelines) (04/02/2021)

https://www.ldc.upenn.edu/data-management/providing-data/ documentation-guidelines (Documentation Guidelines) (04/02/2021)
https://www.ldc.upenn.edu/sites/default/files/ldc-related-works-schema.pdf (LDC Related Works Schema) (04/02/2021)
http://www.language-archives.org/OLAC/metadata.html (OLAC Metadata) (07/15/2021)

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

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 4

Response:

LDC's IT infrastructure, supported by the University's framework, provides highly available storage, backup and disaster recovery for archival data. Workflows exist for each step of the archival storage process. Data is stored on volumes only accessible within LDC and write permissions are only granted to the few staff directly involved with maintaining those volumes. The volumes undergo ongoing back-ups and integrity checks. Physical copies of each corpus produced are stored onsite and offsite. In the event of data corruption, data can be restored from backups or from those physical copies. Additional physical onsite and offsite copies of data are made when any updates are made to the in-house volumes.

Storage processes and procedures are documented on an internal wikis accessible to LDC staff. LDC IT staff manages updates to these documents in area such as volume types, backups, and integrity checks. Publications staff manages updates to documented procedures in areas such as file directory standards, update procedures, and physical duplicates.
URL: https://www.ldc.upenn.edu/about/facilities/it-networking (IT Infrastructure) (04/02/2021)
https://www.ldc.upenn.edu/data-management/curation-distribution (Curation and Distribution Services) (04/02/2021)
https://www.ldc.upenn.edu/data-management (Data Management) (09/09/2021)
https://www.ldc.upenn.edu/data-management/preserving (Preserving Data) (09/20/2021)

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

Reviewer 2

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

Response:

LDC is committed to the long-term accessibility of all data in the LDC Catalog. Every corpus deposited at LDC remains available, including those created in the 1980s and placed at LDC at its founding. There is no difference in preservation levels for published corpora. Redundant backups, multiple drives, and off-site storage of physical copies of all corpora

assure long-term preservation. LDC ensures that data is migrated to new formats, platforms and storage media as required by best practices in the digital preservation community, such as in the risk of obsolescence. Updates are made in such a way as to ensure previous versions are retrievable. Storage provides ongoing, built-in integrity checking to ensure fixity. See “Preserving Data” for more information.
To enable all of the above, providers are required to submit appropriate metadata and submissions data alongside their resource. All providers must sign a distribution agreement with LDC that gives the Consortium the right to store and distribute the submitted resource. LDC has the right to copy, transform, store, and provide access to the data in its catalog and treats ensuring long-term access and understandability as a critical function.

URL :
https://www.ldc.upenn.edu/about/facilities/it-networking (IT Infrastructure) (04/02/2021)
https://www.ldc.upenn.edu/data-management/curation-distribution (Curation and Distribution Services) (04/02/2021)
https://www.ldc.upenn.edu/data-management/providing/publication-process (Publication Process) (04/02/2021)
https://submissions.ldc.upenn.edu/ (LDC Submissions) (04/02/2021)
https://www.ldc.upenn.edu/data-management/preserving (Preserving Data) (09/20/2021)

Reviewer Entry

Reviewer 1

Comments:

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 4

Response:

LDC's catalog entries provide descriptive metadata following Dublin Core and OLAC standards. Resource documentation, including the catalog description, provides structural metadata (how to process/use the data set.) LDC requires data submitters to adhere to technical and documentation standards in order to have their data published with LDC. In the event these standards are not met initially, LDC works with the provider to align their data with these standards. License information provides administrative metadata covering intellectual property, conditions on use and so on. LDC conducts periodic user surveys that provide the community with the opportunity to comment on LDC publications. LDC provides citations to related works as appropriate. LDC staff include leading researchers in many of the disciplines that LDC supports. These researchers rely upon LDC data and provide input into data and metadata quality and adequacy for intended purposes.
URLs: https://submissions.ldc.upenn.edu/ (LDC Submission) (07/15/2021)
https://www.ldc.upenn.edu/data-management/providing/documentation-guidelines (Documentation Guidelines) (04/02/2021)
https://www.ldc.upenn.edu/data-management/providing-data/technical-guidelines (Technical Documentation) (04/02/2021)
https://www.ldc.upenn.edu/about/staff (Staff) (04/02/2021)
https://www.ldc.upenn.edu/data-management/providing/publication-process (Publication Process)

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 4

Response:

LDC maintains workflows from submissions to publication to archiving and curation. These workflows, which maintain the integrity of the data and ensure proper curation practices, and are outlined on LDC's web site. Broadly, our workflow follows OAIS's functional model of 1) Ingest with initial contact with providers where details of the deposit of data are discussed and resolved; 2) Archive Storage & Data Management where quantitative and qualitative checks are carried out on the data on our servers before and after selection. 3) Administrative where agreements are negotiated and data may or may not be selected due to quality, licensing, or mission scope reasons, which is communicated to the provider; 4) Preservation Planning in which data is stored immediately on volumes featuring robust backup and integrity checking; and 5) Access in which data is disseminated according to our distribution policies via download or media. An overall diagram and process is presented on the Publication Process page (https://www.ldc.upenn.edu/data-management/providing/publication-process)
More specific steps, instructions, flowcharts and procedures are maintained on an internal wiki. These procedures are updated routinely by the Publications department to address specific circumstances that arise or changes in technology. Significant updates are reviewed and discussed with the External Relations Manager before implementation.

URLs:
https://www.ldc.upenn.edu/data-management/providing/publication-process (Publication Process) (07/22/2021)
https://www.ldc.upenn.edu/data-management/curation-distribution (Curation and Distribution Services) (04/02/2021)

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 4

Response:

LDC's online catalog provides search capabilities under various criteria, including language, title, author, data source and project, as well as keyword searching in the corpus description. All LDC data have four unique identifiers – DOI, ISBN, LDC identifier and ISLRN (International Standard Language Resource Number). All LDC data have permanent URLs to their individual catalog records that allow users to access public documentation and metadata, which in turn can be used to evaluate the data. In the event of migration to another system, redirects will be set up to ensure any previous URLs still direct users to the appropriate catalog record. This process has already been undertaken once in a catalog redesign in 2014. ISLRN and DOI records are also accessible from persistent URLs. LDC's catalog metadata is automatically harvested daily by Dublin Core-compliant OLAC and displayed with other archives on OLAC's website. LDC provides citation guidelines for its data.
URLs: https://www.ldc.upenn.edu/data-management/providing-data (Providing Data) (04/02/2021)
https://www.ldc.upenn.edu/language-resources/data (Data) (04/02/2021)
https://catalog.ldc.upenn.edu/search (Search the LDC Catalog) (04/02/2021)
https://www.ldc.upenn.edu/data-management/citing (Citing Data) (04/02/2021
http://www.language-archives.org/ (OLAC) (04/02/2021)
https://doi.org/10.35111/wk4f-qt80 (English Gigaword Fifth Edition via DOI) (07/15/2021)

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 4

Response:

The catalog entry provides descriptive metadata following our scheme which is based off OLAC standards, which in turn are based off of Dublin Core. Resource documentation, including the catalog description, provides structural metadata (how to process/use the data set.) LDC requires data submitters to adhere to technical and documentation standards in order for their data to be published with LDC. In the event these standards are not met initially, LDC works with the provider to align their data with these standards. License information provides administrative metadata covering intellectual property, conditions on use and so on. Catalog descriptions and metadata are designed to provide users with sufficient information so that the data can be used for its intended purpose.
Data is provided in formats commonly used by LDC's community. In brief, these are: text in UTF-8 encoding and markup as XML; audio as FLAC-compressed MS-WAV; and video as AVI or MP4. LDC may make exceptions in specific cases when working with providers, and adapts to changing formats when needed. For example, UTF-8 is now the standard encoding for text; it replaced earlier encoding schemes. LDC migrates data to new formats as needed, by for example, providing FLAC speech files for a corpus originally released in sphere format.

URLs: https://submissions.ldc.upenn.edu/ (LDC Submissions) (04/02/2021)
https://www.ldc.upenn.edu/data-management/providing-data/technical-guidelines (Technical Guidelines) (07/15/2021)
https://www.ldc.upenn.edu/data-management/providing/documentation-guidelines (Documentation Guidelines)

(07/15/2021)

https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/share-data-2019-update.pdf (Share Data through LDC) (07/15/2021)

http://www.language-archives.org/OLAC/metadata.html (OLAC Metadata) (07/22/2021)

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 4

Response:

LDC's catalog backend is a MySQL/MariaDB relational database that connects data sets, their attributes and user information. It is enhanced with e-commerce modules that give users control over their user accounts and the ability to join LDC and license data online. This system was built using Ruby, Ruby on Rails and Spree, the Ruby e-commerce platform. Corpora are stored on ZFS file systems with RAIDZ2 and hosted on two independent, mirrored, systems, physically located in two different buildings. For disaster recovery, backups are offsite and geographically dispersed. ZFS features protection against data corruption with regular integrity checking and automatic repair as well as flexible scaling and robust permissions restrictions over sharing protocols.

LDC meets the OAIS functional model:
(1) Ingest: Workflow that includes submissions form from data provider, LDC staff review, data added to holdings.
(2) Archival Storage: Documented process for managing data storage; explicit archiving workflows across data life cycle.
(3) Data Management: Catalog tracks descriptive, technical and structural metadata.
(4) Administration: Staff ensures that access rights and licenses are recorded and tracked.
(5) Preservation Planning: Workflows and procedures in place to preserve data integrity and accessibility.
(6) Access: Data accessible through catalog interface and via communications with LDC's Membership Office.
University and LDC IT staff ensure around-the-clock connectivity for LDC infrastructure which is built upon server grade commercial hardware and standard, open source software. LDC relies on the University's Information Systems & Computing department's service level agreements for provided services, which are reviewed annually. Independently, LDC runs a real-time monitoring system, InterMapper, to ensure the health of the technical infrastructure in all areas. A redundant InterMapper system runs off-site in the event of local issues or interruption.

URL: https://www.ldc.upenn.edu/data-management/curation-distribution (Curation and Distribution Services) (04/02/2021)
URL: https://www.ldc.upenn.edu/about/facilities/it-networking (IT Infrastructure) (04/02/2021)
URL: https://www.ldc.upenn.edu/members/manage-account (Managing Your LDC Account) 04/02/2021)
https://www.isc.upenn.edu/slas (Service Level Agreements) (09/15/2021)

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 4

Response:

LDC is located within an access-controlled building with a guard at the building entrance and card swipes at the building and suite entrances. The card swipe at the suite entrance admits only current LDC staff. LDC computers are kept in locked server rooms on a private network that is maintained by the University's Information Systems and Computing group.
LDC data is protected by real-time monitoring, alarming, redundant backups, multiple drives, and off-site storage. Storage includes frequent recurring integrity checks to guarantee fixity. Access policies prevent accidental or unauthorized changes to data from within LDC. Disaster recovery of data is implemented via backups, mirroring and snapshots, taken daily for data in flux or monthly for static data, stored locally and offsite at Amazon. Beyond that, backups are maintained at regular intervals to supply recovery even if data loss is not noticed immediately. In the event that LDC terminates its operations, all Consortium property remains the property of the University of Pennsylvania.
LDC complies with Penn risk management procedures and requirements such as SPIA (https://www.isc.upenn.edu/security/spia). Additionally, LDC implements its own internal risk analysis, based on input and concerns raised in regular and ad hoc security meetings. Multiple meetings are run each month addressing needs of LDC's technical leads and any of general staff while also informing staff of any current security concerns, such as prevalent phishing scams or security holes.

URLs: https://www.ldc.upenn.edu/about/facilities/it-networking (IT Infrastructure) (04/02/2021)
https://www.ldc.upenn.edu/data-management/curation-distribution (Curation and Distribution Services) (04/02/2021)

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

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:

The naming of specific versions of requirements via year (e.g. 2017-2019) has led to some confusion over when that certification expires. We have found that a number of users and staff have assumed that our certification expired in 2019, the final year of the version, rather than three years after acceptance, in 2021.

Reviewer Entry

Reviewer 1

Comments:
On the whole, this is a very good example of an assessment, kept short and concise in description, with references to the detailed documents and explanations on the DLC website.

Reviewer 2

Comments:
Overall, a solid application from an established repository with good practices.