Assessment Information

CoreTrustSeal Requirements 2020–2023

Repository:	LAC - Language Archive Cologne
Website:	https://lac.uni-koeln.de/
Certification period:	18 April 2023 - 17 April 2026
Requirements version:	CoreTrustSeal Requirements 2020-2022
This repository is owned by:	**Data Center for the Humanities, University of Cologne**

CoreTrustSeal Board

W www.coretrustseal.org E info@coretrustseal.org

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background Information

Repository Type

Please provide context for your repository. You can select one or multiple options.

Compliance level:

Not Applicable - 0

Response:

- Domain or subject-based repository

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Description of Repository

Provide a short overview of the repository.

Compliance level:

Not Applicable - 0

Response:

Language Archive Cologne (LAC) [1] is a subject-based repository and provides services for audio-visual data and in particular audio-visual speech recordings as used by language typologists, field linguists, and other type of linguists, as well as anthropologists, musicologist, historians and other researchers with audio-visual (speech) data in the humanities and social sciences. The Language Archive Cologne provides enhanced curation services for audio-visual data, metadata, time-aligned annotations and related data types.
References:
[1] https://lac.uni-koeln.de

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Designated Community

Provide a clear definition of the Designated Community

Compliance level:

Not Applicable - 0

Response:

The Language Archive Cologne (LAC) supports research, learning and teaching with high quality and dependable digital language resources. The LAC facilitates free and open online access to research data. It does this by preserving digital audio and video data, annotations, and other digital language data in the long term, and by promoting and disseminating these resources. Archiving audio-visual speech recordings together with accompanying annotations as created in language documentation efforts and similar projects are the main focus of the LAC. The designated community includes linguists, especially those working in typology or with interests for a certain language or language family, but also more general researchers with a focus on anthropology, musicology, oral history. Thus our format whitelist and curation efforts are mainly informed by best practices in the field of language documentation, the Department of Linguistics at the University of Cologne provides expertise on this. A special attention is paid to correct and useful information and representation of the language recorded and the geographical location as this will be the first point of entry for most users.
The LAC promotes open and sustainable practices in the use of digital language data in the humanities and provides technical advice to the research community, and supports the deployment of digital technologies in research and teaching. It offers archiving services for audio-visual speech data to researchers – including students and native speakers.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Level of Curation

Select all relevant types of curation.

- Content distributed as deposited
- Basic curation – e.g., brief checking, addition of basic metadata or documentation
- Enhanced curation – e.g., conversion to new formats, enhancement of documentation
- Data-level curation – as above, but with additional editing of deposited data for accuracy

Compliance level:

Not Applicable - 0

Response:

- A. Content distributed as deposited
- B. Basic curation – e.g. brief checking; addition of basic metadata or documentation

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Level of Curation - explanation

Please add the description for your Level(s) of Curation.

Compliance level:

Not Applicable - 0

Response:

The LAC repository itself carries out mainly B level curation. At the minimum this entails the creation of metadata in CMDI format describing the resource as well as ensuring availability of data in appropriate formats.
The data objects are archived as they are provided by the depositor. The repository makes the original deposited objects available in an unmodified way. No format conversion is performed when creating the AIP from the SIP. Metadata undergo a curation process during ingest, where they are enhanced with normdata for fields relating to the language (added glottocodes [1]), the location (added geolocation data) and the actors (added orcid or isni identifiers [2]) and checked for consistency as further described in R12.

The LAC maintains a whitelist of accepted data types and file formats and will only accept deposits in these formats.[3] The archive may support depositors in converting non-whitelisted data formats into accepted formats. However, the LAC does only guarantee support beyond B level curation for a very limited set of file formats on the whitelist (see [4]). In particular, LAC does not convert all data formats to new formats as is characteristic for general C level curation (see also R7 and R8).

References:
[1] https://glottolog.org/
[2] https://orcid.org/, https://isni.org/
[3] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/format-whitelist
[4] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/preservation-plan

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Insource/Outsource Partners

If applicable, please list them.

Compliance level:

Not Applicable - 0

Response:

There are no outsourced partners involved in the operation of the LAC.
LAC is jointly operated by the Data Center for the Humanities (DCH)[1] and the Department of Linguistics (Institut für Linguistik, IfL)[2] in close collaboration and with the support of the Regional Computing Center (RRZK).[3] All partners are part of the University of Cologne (UoC). All staff members are employees of the University of Cologne (UoC) and all technical resources necessary for the operation of the archive and repository are provided within the UoC. The departments and units at the University of Cologne provide different expertise and perform different roles in the operation as further elaborated under “Other Relevant Information”.
References:
[1] http://dch.phil-fak.uni-koeln.de/
[2] http://ifl.phil-fak.uni-koeln.de/?L=1
[3] https://rrzk.uni-koeln.de/

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Significant Changes

Summary of Significant Changes Since Last Application if applicable.

Compliance level:

Not Applicable - 0

Response:

-

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Other Relevant Information

You may provide other relevant information that is not covered by the requirements.

Compliance level:

Not Applicable - 0

Response:

LAC is jointly operated by the DCH and IfL in close collaboration and with the support of the Computing Center RRZK. All partners are part of the University of Cologne (UoC). Each of them perform dedicated roles in the context of the LAC.

The DCH is an institutional part of the faculty of arts and humanities and is responsible for the archiving and data management of research data at the faculty. It steers operations and management of the LAC. The DCH has staff trained in research data management with special attention to arts and humanities. Its primary role in the LAC is technical coordination, data management, general management and networking/interfacing with national and European research infrastructures. The DCH is scientifically guided by its board of advisers and headed by the board of directors, consisting of three professors from the field of digital humanities.[1]

Additionally, the DCH is a CLARIN-B Centre and forms part of the CLARIN ERIC and CLARIN-D infrastructure. The DCH adopts and implements standards and procedures established by the European research infrastructure and observes its expert recommendations. The DCH provides resources and services to the CLARIN infrastructure via the LAC in a CLARIN-compliant way.

The IfL is a research institution renown for its focus and formative influence on language documentation and research on small and under-researched languages. It is currently the only linguistics department in Germany with two full professors for general linguistics. The IfL has methodical and technical expertise in data collection and corpus creation and contributes to the LAC with research data curation, data management, metadata compilation and general scientific guidance.

The RRZK is the computing centre of the University of Cologne – one of the biggest universities in Germany. The RRZK has broad expertise in providing storage and computing capabilities for research and teaching as well as software development in scientific projects. Several departments and service layers of the computing center RRZK are relevant to the LAC. These departments are part of the regular operations of the computing center ranging from hardware maintenance, virtualisation and deployment to software development The particular services, their role in the operation of LAC, and their respective contacts are listed in detail in the internal documentation.[2] Most notably the RRZK contributes to the LAC by providing and managing storage and archiving capabilities and by operating the repository software.

The collaboration of DCH, IfL, and RRZK was strengthened in the project KA^3.[3] The project was funded by the BMBF[4] from October 2016 to September 2019 and received funding for a second phase from October 2018 to September 2020. The main goal of KA^3 was to develop and establish the underlying infrastructure, repository software, and archive procedures.

The activities of the LAC are firmly embedded in a larger context of competence bundling in the field of digital humanities at the UoC and faculty of arts and humanities, which is also home to the Department for Digital Humanities[5] and the Cologne Center for eHumanities.[6]

See R5 for more information on funding and staff and R6 for more information on expert guidance and governance.

The LAC repository is part of the CLARIN B Centre in Cologne and therefore a part of CLARIN-D (Common Language Resources and Technology Infrastructure Deutschland) - a web and centres-based research infrastructure for the social sciences and humanities. The aim of CLARIN-D and its service centres is to provide language data, tools and services in an integrated, interoperable, and scalable infrastructure for the social sciences and humanities. The research infrastructure is rolled out in close collaboration with expert scholars in the humanities and social sciences, to ensure that it meets the needs of users in a systematic and easily accessible way. CLARIN-D is funded by the German Federal Ministry for Education and Research. CLARIN-D collaborates with and uses infrastructural components of the European CLARIN[7] initiative. Research standards to be met by the CLARIN services centres, technical standards and solutions for key functions, a set of requirements which participating centers must provide, as well as plans for the sustainable provision of tools and data as well as their long-term archiving have been developed. DCH is a certified centre of type B. CLARIN distinguishes a number of different centre types that have different impact for the language resources and tools infrastructure. Type B centres offer services that include the access to the resources stored by them and tools deployed at the centre via specified and CLARIN compliant interfaces in a stable and persistent way. Within CLARIN-D the following requirements hold for centres of type B[8] and are fulfilled by DCH: Centres need to offer useful services to the CLARIN community and to agree with the basic CLARIN principles (own architecture choice, explicit statement about quality of service, usage of persistent identifiers, adherence to agreed formats, protocols and APIs). Centres need to adhere to the security guidelines, i.e. the servers need to have accepted certificates. Centres need to join the national identity federation where available and join the CLARIN service provider federation to support single identity and single sign-on operation based on SAML 2.0 and trust declarations. In case all resources at a centre are open, setting up a Service Provider is optional. Centres need to have a proper and clearly specified repository system and participate in a quality assessment procedure as proposed by the Data Seal of Approval or MOIMS-RAC approaches. Centres need to offer component based metadata (CMDI) that make use of elements from accepted registries such as CLARIN Concept Registry in accordance with the CLARIN agreements, i.e. metadata needs to be harvestable via OAI PMH. Centres need to associate PIDs records according to the CLARIN agreements with their objects and add them to the metadata record. Each centre needs to make clear statements about their policy of offering data and services and their treatment of IPR (intellectual property rights) issues. Each centre needs to make explicit statements to the CLARIN boards about its technological and funding support state and its perspectives in these respects. Centres need to employ activities to relate their role in CLARIN to the research community in order to guarantee a research based status of the infrastructure and allow researchers to embed their services in their daily research work. Centres that are offering infrastructure type of services need to specify their services for CLARIN and the terms of giving service. Centres are advised to participate in the Federated Content Search with their collections by providing an SRU/CQL Endpoint. This content search is especially suitable for textual transcriptions and resources. A short overview of all requirements for centres of type B is also given in the form of a checklist. [9]

The CLARIN mission is to create an infrastructure that makes language resources and technology available and readily usable to scholars of all disciplines, in particular the humanities and social sciences. In our age we are presented by many challenges as we deal with language in electronic formats, in spoken, written, and multimodal forms, as a carrier of information, as an object of study, and otherwise. The volume of texts and recorded spoken texts is enormous, and it is growing exponentially. The sheer size of this material makes the use of computer-aided methods indispensable for many scholars in the humanities and in neighbouring areas who are concerned with language material. CLARIN is committed to boosting humanities research in a multicultural and multilingual Europe, by facilitating access to language resources and technology for researchers and scholars across a wide spectrum of domains in the humanities and social sciences (Krauwer, 2008).

The LAC is a member of the DELAMAN network of archives dedicated to endangered languages and music. [10]

The LAC is part of the German National Research Data Infrastructure (NFDI). [11] In the NFDI Consortium Text+ [12], the LAC is a data and competence center in the Data Domain “Collections”. [13]

References:

[1] https://dch.phil-fak.uni-koeln.de/ueber-das-dch/satzung-und-beirat
[2] https://redmine.uni-koeln.de/projects/ka3/wiki/Akteure_und_Rollen (access protected, pdf version attached)
[3] https://dch.phil-fak.uni-koeln.de/forschung/ka3-koelner-zentrum-analyse-und-archivierung-von-av-daten
[4] Federal Ministry of Education and Research – BMBF, https://www.bmbf.de/
[5] https://dh.phil-fak.uni-koeln.de/
[6] http://cceh.uni-koeln.de/
[7] https://www.clarin.eu/
[8] https://www.clarin.eu/node/3542
[9] https://www.clarin.eu/content/checklist-clarin-b-centres
[10] https://www.delaman.org/
[11] https://www.nfdi.de/?lang=en
[12] https://www.text-plus.org/
[13] https://www.text-plus.org/en/research-data/data-and-competence-centres/

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Organizational Infrastructure

R1 Mission/Scope

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The Language Archive Cologne has an explicit mission to provide access to and preserve audio-visual (speech) data. The mission statement[1] of the Language Archive Cologne states:
The Language Archive Cologne (LAC) supports research, learning and teaching with high quality and dependable digital language resources. The LAC facilitates free and open online access to research data. It does this by preserving digital audio and video data, annotations, and other digital language data in the long term, and by promoting and disseminating these resources. The LAC promotes open and sustainable practices in the use of digital language data in the humanities and provides technical advice to the research community and supports the deployment of digital technologies in research and teaching.
The LAC with its mission is integral part of the Data Center for the Humanities (DCH) of the Faculty of Arts and Humanities at the University of Cologne. The DCH has the task to secure and provide long-term preservation for diverse types of digital data and make them accessible.
The LAC repository is serving as the repository for audio-visual data of the CLARIN-D resource centre DCH. The mission of CLARIN-D is to provide “linguistic data, tools and services in an integrated, interoperable and scalable infrastructure for the social sciences and humanities“.[2] Therefore, a repository in which data, tools and accompanying metadata are archived on a long term basis has to be operated by such a resource centre.
References:
[1] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/mission-statement
[2] https://www.clarin-d.net/en/

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Compliance Level Comment: ok

R2 Licenses

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The Language Archive Cologne supports open access and encourages the use of open licences, in particular licenses of the Creative Commons license family, in the humanities. However, we recognize that audiovisual recordings may contain sensitive personal or cultural data as well as performances that may require protection and access restrictions. In individual cases, the usage restrictions imposed by copyright law may be required to protect the rights of consultants and performers. Further access restrictions can be necessary to protect personality right, traditional knowledge, or other goods requiring protection. Restricted access to specific data is currently being implemented. While the LAC currently only disseminates publicly accessible resources, this access control system will allow enforcement of access restrictions such as requiring registration or granting access upon individual request.

Mutual rights and obligations of the depositors and the repository are being governed by the depositor agreement[1] which has to be signed by both parties prior to depositing data in the repository. The contract entails regulation of access conditions, the terms under which the repository may make the deposited data available to third parties, as well as an explicit licence selection.

The LAC encourages the publication of research data Creative Commons Attribution 4.0 International (CC BY 4.0) and Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). Currently, the LAC also contains data that is either published under standard copyright or licensed under the Creative Commons Attribution-NonCommerical-ShareAlike 4.0 International Public License (CC BY-NC-SA 4.0). Several data sets are available under the restrictions of the DoBeS code of conduct.[2]

The archive monitors data consumer compliance with licenses and any further specified restrictions. Access restrictions are technically enforced and the efficiency of these measures is regularly tested. The LAC has a help desk (lac-helpdesk@uni-koeln.de) that allows affected parties as well as others to call data misuse and violations of licenses to our attention. In the case of non-compliance with licenses and any further specified restrictions, the archive will take appropriate action. All licenses and access restrictions (will) conjunct to disciplinary and ethical norms (see also section R4).

Besides monitoring the ethical and legal compliance of users (consumers), the archive also monitors the ethical and legal obligations of depositors. The archive is approachable by consultants, speech communities and their representatives. In case of non-compliance, the archive will take appropriate action. This may include take-down of problematic materials and in the case of repeated offences or particularly strong offences revocation of the depositors account.

The LAC discourages the publication of audiovisual recordings and data that may have adverse consequences for involved parties. The LAC offers special archiving services for some types of sensitive data (see R4).

Access to the LAC repository is governed by its terms of use[3], which details terms of service, privacy policy, and regulations for data deposit and data access.

References:
[1] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/depositor-agreement
[2] http://dobes.mpi.nl/ethical_legal_aspects/DOBES-coc-v2.pdf
[3] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/terms-of-use

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Compliance Level Comment: ok

R3 Continuity of access

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The LAC guarantees 10 years preservation of research data to researchers and funding agencies. However, the mission scope aims for an indeterminate preservation of the data. Especially the recordings of highly endangered languages constitute documents of cultural heritage and every measure is taken to preserve these records for ensuing ages. The LAC ensures ongoing access to and preservation of the data through several measures. In particular, the LAC is situated in an institutional setting that ensures long-term stability:

The LAC is hosted by the computing centre of the University of Cologne (RRZK) which is the central institution for technical infrastructure at the University of Cologne. The computing centre guarantees permanent technical support to the LAC.

Backup strategies as described in section R9 make sure, that there is an ongoing access to and preservation of the objects hosted by the repository. The repository will not allow any deposit of data without a signed agreement specifying the handling of the data and access to it in detail.

On the organizational side, the Language Archive Cologne is part of the Data Center for the Humanities (DCH) of the University of Cologne. The DCH is an institutional part of the Faculty of Arts and Humanities at the University of Cologne and directly associated with the dean of the Faculty of Arts and Humanities.

The Faculty of Arts and Humanities at the University of Cologne is the legal successor of the DCH. This is also stated in the by-laws of the DCH.[1] If the faculty would decide to close down the Data Center for the Humanities, the faculty has the legal obligation to assure the long-term preservation of all projects, infrastructures, and systems managed by the DCH.

Beyond the local institutional setup, the repository is part of the CLARIN B-centre in Cologne. CLARIN centres commit to ensuring long-term availability, access and to preservation of data sets submitted to their repositories, as set out in their Mission statements. CLARIN centres are setup as a distributed network, where each centre institution is a hub of the digital humanities and brings its own financial resources into CLARIN-D, which ensures continued availability. Additionally, in case of a withdrawal of funding the repositories content would be transferred to another CLARIN centre as formulated in a Memorandum of Understanding by the centres of CLARIN-D.[2] The legal aspects of the process of relocating data to another institution is addressed by templates of license agreements provided in CLARIN-D.

References:
[1] http://dch.phil-fak.uni-koeln.de/sites/dch/user_upload/Satzung_DCH__11.07.2018_.pdf (section 4)
[2] https://www.clarin-d.net/en/about/centres/mou-taking-other-centre-s-data

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Compliance Level Comment: ok

R4 Confidentiality/Ethics

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The LAC follows the conventions of CLARIN: Metadata is always accessible, but the original data may have restricted access. Moreover, the archive aims at making research data as accessible as possible to support scientific collaboration. Depositors must sign a depositor agreement[1] and state that they own all necessary rights required to deposit the data. Furthermore, they are obliged to adhere to general privacy and intellectual property rights protocols. In particular, data must be deposited anonymized when appropriate. The LAC requires data consumers to comply with the DFG code of conduct for good scientific practice[2] and must confirm that they will use resources only in the intended way.

However in individual cases protection of personality, privacy rights, as well as indigenous intellectual property (traditional knowledge), compliance with national law or local norms of the speech community, as well as embargo arrangements, may preclude disclosure and free circulation of the data under open licences. This must be specifically observed in audio-visual language data which typically depicts consultants and performers in human interaction. The LAC aims at mitigating conflicts of interest and misconceptions by consulting depositors in early stages of project planning. The LAC and DCH inform researchers in their consulting workflows about implications of licensing and disclosure, and encourage researchers to incorporate resolvement of legal questions in their project plan and to seek informed consent from the consultants where applicable.

When preparing a deposit, the data managers and prospective depositors will discuss the depositor agreement, address licencing conditions and carry out a risk assessment regarding disclosure of the data in question in an admission consultation. The data managers will further review the data and make an independent assessment of the risks.

If the result of this evaluation is, that the data should not be openly accessible, only the metadata is published in the LAC repository to ensure discovery and citability of the datasets. The primary data itself will not be stored on web-accessible repository systems but instead archived on TSM long-term storage facilities (see R9 and internal documentation[3] for further information). The technical interfaces to the long-term storage are protected and are only accessible through dedicated machines within the university network. The risk of accidental exposure of the data through the internet can be considered as extremely low.

Repository users may contact the depositor through the LAC data managers and request access to the data. The depositor as the rights-holder decides whether the requesting party may be granted access to the data and under which conditions and instruct the LAC to act on his or her behalf.

Guidelines[4] and a depositor agreement[5] are provided for depositors. A data user agreement[6] is provided for data consumers. Furthermore, the LAC offers advisory services and training for depositors and data consumers that also covers licensing and disclosure risk assessment. This activity is organized in the context of the CLARIN Knowledge-Centre for linguistic diversity and language documentation (CKLD)[7], which the Language Archive

Cologne operates together with the Endangered Language Archive (SOAS, London), the HZSK (Hamburg) and the ZAS (Berlin). The CKLD supports researchers from the early planning phase to the realization of language documentation and other fieldwork based research projects as well as typological research. It provides information and assistance relating to fieldwork, archiving and data-related methodological aspects (see R6 for more information on the role of CKLD).

References:
[1] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/depositor-agreement
[2] http://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/gwp/
[3] https://redmine.uni-koeln.de/projects/ka3/wiki/Prozesse_im_Datenmanagement (access protected, pdf attached)
[4] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/depositor-guidelines
[5] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/depositor-agreement
[6] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/data-user-agreement
[7] http://ckld.uni-koeln.de/

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Compliance Level Comment: ok

R5 Organizational infrastructure

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:

The guideline has been fully implemented in the repository - 4

Response:

The LAC is permanently funded by the University of Cologne. It is part of the basic service portfolio of the Data Center for the Humanities (DCH) and jointly operated by the DCH and the Department of Linguistics (IfL) of the University of Cologne. It is technically operated by the computing centre of the University (RRZK).

The staff of the LAC consists of institutionally funded permanent employees and project funded employees and assistants. The permanent institutional funding at DCH, IfL, and RRZK covers technical infrastructure, travel and other expenses, and core operations, including data curation, ingest and preservation, technical operation and maintenance, as well as administration, user support and management. The core operations are carried out by 1.25 FTE at the DCH and IfL as well as by staff at the computing center. The positions are

- two permanent 0.5 FTE positions at the DCH trained in research data management with special attention to the arts and humanities These two positions contributr management, technical coordination, data management, as well as networking/interfacing with national and European research infrastructures.[1]

- The IfL has permanent staff with expertise in data collection, curation as well as metadata compilation and provides guidance regarding decision of scientific and discipline-specific relevance. At least one 0.25 FTE position is permanently dedicated to the operation of the archive.

- Several departments and service layers of RRZK which are part of the regular operations of the computing center ranging from hardware maintenance to virtualisation and monitoring ensure the continuous technical operation of the LAC. The particular services, their role in the operation of LAC and their respective contacts are listed in the internal documentation which you will find attached to this application.[2]

The two 0.5 FTE positions at the DCH have duties beyond the management and operation of the LAC. However, the LAC is fully integrated into the service portfolio as well as the consultation, data management, and data archiving workflows at the DCH. As a result, the initial depositor support and general user support are part of the general consultation workflow of the DCH. Budget planning and project development are also part of the general management of the DCH.

Larger or more work intensive deposits are supported by project funded LAC staff. These positions are funded through the data handling budget of the project or by directly including a position at LAC in the project application. The development of new technical features and larger efforts in the advancement of policies and workflows are funded via project applications. As an example, the LAC is currently part of a large project which aims at developing quality criteria for annotated audio-visual speech corpora. [3]

The LAC is part of the German National Research Data Infrastucture (NFDI). [4] For contributing its services to the NFDI consortium Text+ [5], the LAC receives an additional 0.5 FTE position from the infrastructure consortium (currently funded September 2021–2026). [6]

The DCH as an institution or the employees in the two permanent positions at the DCH are members of DINI [7], DELAMAN [8], RDA, as well as several academic and professional societies (including the German Linguistic Society, the German Endangered Language Society, and the German, European, and International Digital Humanities society).

The LAC emphasizes the close integration of its services into the research and teaching activities at the Univesity of Cologne. This integration in the teaching an research as well as the membership of the LAC and DCH in several professional networks and infrastructures provides rich and up-to-date training and development oportunities for LAC staff. Some project funded positions are explicitly academic qualification positions.

References:

[1] <https://dch.phil-fak.uni-koeln.de/ueber-das-dch/team>.
[2] <https://redmine.uni-koeln.de/projects/ka3/wiki/Akteure_und_Rollen> (access protected, attached as pdf)
[3] <https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html>
[4] <https://www.nfdi.de/?lang=en>
[5] <https://www.text-plus.org/>
[6] <https://www.text-plus.org/en/research-data/data-and-competence-centres/>
[7] <https://dini.de/dini/mitglieder/mitgliederliste/institutionen/>
[8] <https://www.delaman.org/members/language-archive-cologne/>

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Compliance Level Comment: ok

R6 Expert guidance

The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

LAC is jointly operated by the DCH and IfL in close collaboration and with the support of the Computing Center RRZK (see also section R0 “Other relevant information”).
The DCH is an institutional part of the faculty and is the centerpiece of the faculty’s strategy to provide organized research data management and archiving for research activities in the arts and humanities. It steers operations and management of the LAC. With the adoption of the by-laws of the DCH by the faculty council in July 2018, a clear system of governance and scientific guidance has been implemented.[1] The by-laws provide that the DCH is headed by a board of directors[2], consisting of a speaker and two further members, and is guided by a board of advisers[3]. The current directors hold professorships in the field of digital humanities at the faculty. The current speaker is also director of the department for digital humanities. The board of advisers consists of representatives from departments and institutions – internal and external to the UoC – that are fundamental to the operational and

strategic development. The representation of the Regional Computing Center RRZK and the University Library in the board of advisers is mandated by the by-laws. Other members include the director of the Department for Linguistics (IfL) and a representative from the research data management initiative of North Rhine Westphalia[4]. These structural instruments ensure that operational and strategic decisions are informed by newest developments in research data management and digital humanities, as well as institutional and supra-regional polities.

Scientific guidance and feedback from the scientific community is indispensable for the LAC in order to carry out its mission as a subject-based repository. It is mainly incorporated through the involvement of the Department of Linguistics. It is currently the only linguistics department in Germany with two full professors for general linguistics. The Department of Linguistics has staff from project and permanent budget dedicated to the operation of the LAC (see also section R5).

The CLARIN Knowledge Centre CKLD[5] provides an important forum for expert exchange on archiving and curation of audio-visual data in language research. CKLD is an association of major European research data and competence centres engaged in linguistic diversity research and language documentation. CKLD actively works towards harmonisation of methods and workflows across the centres and collectively offers training materials and activities for researchers.

Its current members are:

- Data Center for the Humanities (DCH) at the University of Cologne
- Department for Linguistics (IfL) at the University of Cologne
- Hamburg Centre for Language Corpora (HZSK) at the University of Hamburg
- Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages (INEL) at the Academy of Sciences and Humanities in Hamburg
- Endangered Languages Archive (ELAR) at SOAS University of London
- SOAS World Languages Institute at (SWLI) SOAS University of London
- Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS), an Leibniz Institute based in Berlin

Beyond these activities, DCH maintains relationships and is a member of various relevant organisations in the field of research data management in the humanities[6]. Most notably it is a CLARIN-B Centre and adopts and implements standards and procedures established by the European research infrastructure and observes its expert recommendations.

CLARIN-D is supported by external advisory committees. The International Advisory Board (IAB), CLARIN-D’s scientific advisory board, is a group of CLARIN-D external experts who are consulted on new developments and discuss strategic and content related developments, also with a bird-eye view of other developments in the communities. With experienced experts from various backgrounds, a high-profile international committee was formed for this purpose. Members of the IAB are currently: Helen Aristar-Dry, Christiane Fellbaum, Björn Granström, Helge Kahler, Jan Christoph Meister, John Nerbonne, and Achim Streit. The joint Technical Advisory Board (TAB) of CLARIN-D and DARIAH-DE is a committee that supports collaboration on the fundamental technical level between two large research infrastructures for the humanities and social sciences. The issues of the Collaboration are: questions of technical protocols, infrastructural requirements on the level of archiving, interconnection, search, etc. Based on requirements, small working groups (for example on persistent identifiers, authorization and identification) are being formed in areas with an overlap of requirements. This avoids duplication of developments and allows an increased efficiency in implementation, but also interoperability where overlaps exist. This includes for example an option to grant access to one infrastructure for users of the other. Members of the Technical Advisory Board are currently: Jonas Beskow (University of Stockholm), Carol Goble (University of Manchester), Jan Hajic (Head of the Prague CLARIN Centre), Ed Hovy (University of Southern California), Michael Lautenschlager (German Research Centre for Geosciences, Potsdam), Gerhard Schneider (University of Freiburg), Toma Tasovac (Digital Humanities Centre, Belgrade), Melissa Terras (University College London) and Claire Warwick (University College London). The TAB is currently restructured and its new composition will be announced soon.

CLARIN is committed to boosting humanities research in a multicultural and multilingual Europe, by facilitating access to language resources and technology for researchers and scholars across a wide spectrum of domains in the humanities and social sciences (HSS). To reach this goal and to contribute to overcome the traditional gap between the Humanities and the Language Technology communities we established an active interaction with the research communities in HSS in so called discipline-specific working groups. These groups act as a link between the CLARIN-D resource centres and the research communities which represent the users of the CLARIN-D infrastructure. Currently eight working groups act as consultants for the needs of the humanities, social sciences and particular disciplines. All together they consist of more than 100 academic professionals. Their main role is to advise CLARIN-D during the development and implementation of the infrastructure so that these efforts can best meet the needs of all research communities involved. The working group chairs further coordinate dissemination and best practice using CLARIN-D services in their member communities. CLARIN-D organizes joint activities of the working groups. This includes the organization of working group meetings, organization of specialized and interdisciplinary workshops and the creation of joint reports. Further, communications between CLARIN-D centres and the working groups as well as groups among themselves are coordinated. Virtual meetings are held on a monthly basis. Contents of the curation projects and activities of the WG are published on the CLARIN-D Website[7]. For communication, mailing lists and wiki contents are maintained.

More recently, the DCH has been involved in the German National Research Data Infrastructure (NFDI) [8] and in particular in the consortium Text+[9] as well as one other NFDI Consortium (NFDI4Culture) [10] as well as two applications for further consortia (NFDI4Objects [11] and NFDI4Memory [12])

Lastly, the LAC is part of the DELAMAN Network [13], an "international network of archives of data on linguistic and cultural diversity". Its regular meetings and mailing lists provide insight into international innovation in all areas concerning the archiving of audio-visual language data.

References:

[1] https://dch.phil-fak.uni-koeln.de/ueber-das-dch/satzung-und-beirat
[2] https://dch.phil-fak.uni-koeln.de/ueber-das-dch/team
[3] https://dch.phil-fak.uni-koeln.de/ueber-das-dch/satzung-und-beirat
[4] Landesinitative NFDI der Digitalen Hochschule NRW, https://www.fdm.nrw/.
[5] http://ckld.uni-koeln.de/

[6] https://dch.phil-fak.uni-koeln.de/vernetzung/verbaende-gremien-arbeitsgruppen
[7] https://www.clarin-d.net/en/disciplines/
[8] https://www.nfdi.de/?lang=en
[9] https://www.text-plus.org/
[10] https://nfdi4culture.de/
[11] https://www.nfdi4objects.net/
[12] https://4memory.de/
[13] https://www.delaman.org/

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

I accept this part of your application, but I must admit, that it would be very helpful for you to not only network with other repositories but also with communities for digital preservation as well like e. g. nestor in Germany or other international networks like the dpc or opf.

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Digital Object Management

R7 Data integrity and authenticity

The repository guarantees the integrity and authenticity of the data.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

LAC makes the original deposited objects available in an unmodified way. The repository takes responsibility for the integrity and authenticity of the data and metadata during the processes of ingest, archival storage, and data access. The ingest process ensures the completeness and validity of the data and metadata (see R8 for more detail). Technically, the data are stored as OCFL [1] objects in LAC (with sha512 hashes for verification of data integrity).

However, the repository reserves the right to disseminate or archive the data in suitable alternative formats other than the ones provided by the depositor. This can be relevant when audio-visual data is requested as a media stream by a data consumer. These instances are secondary dissemination formats and do not affect the format or validity of the archived data and the original deposited objects are always available in an unmodified way for download.

The repository would only make changes to the archival object if a file format became obsolete and superseded.[2] Data may also be updated, replaced, enhanced or retracted by the data producer if there is justified cause in accordance with the submission policy of the repository.[3] The archive policies and guidelines are available on the DCH website.[4] The data curators are trained to decide on inquiries based on this policy. In case of changes by the data producer, the original data are kept and the repository creates a new digital object with a new PID, which refers to the previous version via its PID.

The repository only accepts works from the original data producers, who are acknowledged as such in the CMDI metadata and the derivative DC metadata available via the OAI PMH interface.[5] The repository uses CMDI relations to link between bundles and the files they contain as well as between bundles and the collection they are part of. For more technical details on the underlying data modelling see R8 and on the ingest processes which generate these interrelated objects see R12. Deposits are only accepted after a due diligence process involving a check of the identity of depositors and clarification of all legal issues along the lines described in R2 and R4.

On the low infrastructure level file systems used in virtual machines are setup on hard disks in Hardware Redundant Array of Independent Disk System (RAID 5). Therefore it is guaranteed that if a disk crashes, the data is not lost. The data is stripped with a distributed parity for enhanced performance and can be defined very fine-grained. Blocks of data are distributed over all disks in the cluster level implementation, the enables RAID 5 implementations to keep data safe in a duplicated manner, without increasing the disk-usage the same way. Due to the fact that a hardware RAID controller is in use, the functionality of patrol read ensures early read errors on hard disks. These patrol reads are a common way to prevent serious hardware defects at an early stage. Confer R9, R15, and R16 for more detailed descriptions of the technical aspects of integrity and authenticity assurance.

References:
[1] <https://ocfl.io/>
[2] <https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/depositor-agreement>
[3] <https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/depositing-policy>
[4] <https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides>
[5] <https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/depositor-agreement>

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R8 Appraisal

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The LAC is a research data repository for audio-visual speech recordings and immediately connected data types such as annotations – especially time aligned annotations such as transcription and linguistic annotations. Collection policy as well as data and metadata requirements are focused on these purposes. Policies and criteria for depositing speech corpora and other audio-visual resources are published in the LAC Depositing Policy document.[1]

Quality control procedures for metadata and data formats are in place and are applied in the deposit process. Automatic reports are generated and Data sets that do not meet the requirements in regard to metadata format and content or data and file formats are returned to the depositor with guidance to rectify the data set and make it format compliant.

The LAC maintains a whitelist of accepted data types and file formats.[2] Whitelisted formats are monitored for obsolescence. Depositors who wish to archive non-whitelisted data formats are required to contact the archive manager. The LAC will then determine whether the data formats in question will be included into the portfolio of the repository and added to the whitelist or whether the data can be migrated to another whitelisted format.

Following general CLARIN standards, metadata for the LAC repository must be provided in the CMDI format with unique references to the actual resources.

The archive provides recommendations for file formats as part of the LAC submission guidelines.[3] For PCM encoded audio data and video recordings in mov- and mp4-containers the archive offers browser-based players in addition to general file download services.

In case a dataset offered to the LAC should not fall within the mission profile, the depositor will be referred to other more suitable archives by the archive. If the dataset cannot be made available to the public for privacy reasons, the DCH offers dark/cold long-term storage options.

In case a dataset or a bundle is to be removed from the LAC, a flag will be added to the metadata ("Item removed by request of depositor."). The corresponding media files will be replaced by a text-file with the same information. This workflow ensures that the PID's will still resolve and the user is informed about their removal.

References:
[1] <https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/depositing-policy>
[2] <https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/format-whitelist>
[3] <https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/submission-guidelines>

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R9 Documented storage procedures

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

Compliance level:

The repository is in the implementation phase - 3

Response:

All processes and procedures in the managment of archival data are documented in the internal LAC-handbook which is updated regularly. The status of a single dataset is tracked via a GitLab-Issue-Board from drop-off through ingest to dissemination. Information for depositors on our processes can be found in the preservation policy, available under: https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/preservation-plan.
The basic path of a dataset through the process via different storage locations is sketched in Figure 4 under https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/archive-setup: Data can be submitted via Sciebo, a sync and share service for universities in North Rhine-Westphalia or via SOFS, the online storage system of the University of Cologne (this option is only possible for members of the university) or via drop-off of a physical storage medium. The data is then moved to the virtual machine of the Data Center for the Humanities for conversion and quality checks. Then, the data can be staged at the LAC-dev-repository and then published at LAC-prod (both managed by the RRZK). Lastly, a copy of the AIP is send to the Long Term Storage of the RRZK.
The repository software of the LAC runs inside a container-virtualized application cluster, which is hosted on multiple virtual machines at the computing centre of the University (RRZK). Storage and virtualized machine nodes are mainly based on shared Bladecenter implementations that enable virtualization via VMware ESX. Redundant air conditioning, redundant uninterrupted power supplies, early fire detection and fire suppression systems are provided. Accessing the physical servers is limited to authorized staff and the maintenance of the systems is performed by a team of trained personnel. The virtualized machines are secured via specific firewall and network settings. To keep the underlying hardware up to date, the components are upgraded, checked and swapped in regular intervals.
Strategies for backup, recover and multiply are separated in three major areas. This areas become more clear when reading the internal document wiki: https://redmine.uni-koeln.de/projects/ka3/wiki/System-Architektur (access protected, pdf attached)
The areas mentioned above are for explanatory reasons structured into three parts:

High Level (software)
Low Level (virtual machine infrastructure)
Bare Subsistence Level (hardware infrastructure)

1.) The implemented and deployed software is backed up inside a Git repository, hosted by the RRZK. Parts of the repository's data is stored inside a high-available MARIA-DB cluster, which is hosted and secured by the RRZK. The data file – the actual archival objects – and metadata files as well as configuration files and data the application protuces at runtime are stored inside a dedicated network storage (NFS). This NFS is backed up in multiple ways: The NFS backup is scheduled every 2 days and has the current and last recovery provision always available in place. Every day at 11pm an incremental backup stores the relevant folders at the NFS itself at a specific location to backup. Every week this location is backed up via TSM to the tape storage hosted by RRZK. The incremental backup has a retention period of 12 month. The tape backup has a retention period of one year (updated every week) and the NFS internal backup has a retention period of 4 days.
For more information on regarding the TSM Backup, please visit the internal document: https://redmine.uni-koeln.de/projects/ka3/wiki/System-Architektur#TSM (access protected, pdf attached)
2.) The complete virtual machine infrastructure is planned and created via a continuous deployment approach based on container virtualization. This provides two advantages for recovering the software architecture: Everything is recoverable via blueprints of applications containers and this can be achieved in an automated manner. In case of redeployment of virtual machines configurations, an orchestration to configure and administer virtual machines is available.
For more information on regarding the container virtualization, please visit the internal documentation, available under: https://redmine.uni-koeln.de/projects/ka3/wiki/Prozesse_im_Datenmanagement (access protected, pdf attached)

3.) The low-level infrastructure can be separated into two layers: the virtual machines and the storage-layer. Storage for virtual machines is backed up via vdisk images to a vmfs (VmwareFilesystem) storage. This storage is backed every week to the second storage location redundantly. Furthermore, two snapshots of a performing virtual machine are backed up two days past within a nightly snapshot. The storage layer is backed up for disaster recovery on a daily basis to the second storage location at RRZK. For further information read more in sections R15 and R16.
Part of the archiving workflow is the integrity check of the data and the metadata by the archive manager. This is done both manually and automatically. The metadata is validated against an XML Schema and manually evaluated for completeness and soundness. The object data is tested for syntactic correctness if possible. All datastreams and versions are equipped with a sha512 checksum. For further details of the ingest part of the archiving workflow see also R12.

Reviews

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

R10 Preservation plan

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

Compliance level:

The repository is in the implementation phase - 3

Response:

The responsibilities concerning long-term preservation and the management of the LAC are jointly met by the LAC-team proper, consisting of the LAC administration and LAC management, by the computing center of the University of Cologne (RRZK) and by the Data Center for the Humanities (DCH). As illustrated in Figure 3 under https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/archive-setup, the LAC Administration is in charge of technical coordination, preservation planning, strategic planning, engagement with national and international research data infratstructures, user interaction (with depositors and consumers), public relations and certification. The LAC management is responsible for data management, curation, archiving, quality assurance and the conversion of legacy data. The RRZK operates the repository software, provides the server infrastructure and is in charge of its administration, and monitors the repository software. Lastly, the DCH operates the frontend and is the technical contact for SAML-federations.

The data provider retains all intellectual property rights to their data. The depositor must grant distribution rights to the LAC. Enforcing licenses by data users in the case of misuse is conducted by the property rights owner. Crisis management is based on the technical solutions described in R9.

The LAC documents its approach to preservation in the preservation policy.[1] The legal relationship between the depositor and the repository is described in the depositor agreement.[2] Recommendations for depositors about technical standards for submissions can found in the submission guidelines and the format whitelist.[3]
The LAC archives all metadata and data in such a way that they can be easily migrated to and mirrored at other CLARIN resource centers. All metadata and data have a persistent identifier (PID), and are stored as self contained XML files. Legal aspects of the process of relocating data to another institution is addressed by templates of license agreements provided in CLARIN.

The policies in place enhance the chance of future interpretability of the data. The LAC format whitelist (see also R8) makes future conversions to other formats more feasible. As much as possible open (non-proprietary) file formats are used. For textual resources, XML formats are used whenever possible, to make future interpretation of the files possible even if the tool that was used to create them no longer exists. Text is encoded in Unicode (UTF 8) to ensure future interpretability. When a particular file format is in danger of becoming obsolete, appropriate curation steps take place. Additionally, all resources in the repository (metadata and data) are equipped with a checksum.
The current approach on long term archiving is to support bitstream preservation. This is carried out via the well established solution Backup-TSM (IBM) hosted by the RRZK. For additional information about the Backup-TSM solution please see R9. As described, data backup is a fully automated process in LAC. This is documented in the internal documentation.[4] The setup of the automated backup procedures is documented in the internal wiki[5] and the whole backup configuration of the repository and its relation to other parts of the repository is also documented in the internal documentation.[6] These

documents enable the technical staff to monitor backup processes and to intervene, if an error in the execution of the automated backup would occur.

The computing center is continuously evaluating its long term preservation strategies. It does this in cooperation with the local university library. These two institutions have recently evaluated the specific requirements of a long term preservation solution at the University of Cologne. The project in which the evaluation was conducted was lead by the state’s library centre for universities (HBZ). The project had the claim to proof the scalability of a long-term preservation solution in the federal state of North Rhine-Westphalia. The LAC and its KA³ repository was one of these pilot projects.

References:
[1] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/preservation-plan
[2] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/depositor-agreement
[3] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/submission-guidelines,
https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/format-whitelist
[4] https://redmine.uni-koeln.de/projects/ka3/wiki/Automatismen#Backup
[5] https://redmine.uni-koeln.de/projects/ka3/wiki/Prozesse_im_Datenmanagement#Backup (access protected, pdf attached)
[6] https://redmine.uni-koeln.de/projects/ka3/wiki/System-Architektur#TSM (access protected, pdf attached)

Reviews

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

R11 Data quality

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:

The guideline has been fully implemented in the repository - 4

Response:

Data and in particular metadata quality have been a research focus at the DCH in recent years. The LAC is currently part of a large project which aims at developing quality criteria for annotated audio-visual speech corpora and the DCH is responsible for the work package “Metadata”. [1]

The Language Archive Cologne is integrated into the Common Language Resources and Technology Infrastructure (CLARIN), which implements several channels through which members of the designated communities can give feedback on data and metadata hosted by its certified centres. For more details about the data and metadata policies of the LAC, see R8. To ensure data quality workflows are reported in R12. Details about the processes that ensure the integrity and authenticity of the data can be found in section R7.

In order to make sure the deposited datasets are complete and understandable, the LAC conducts an initial sanity check via a python script upon receicving the SIP. In this step, metadata generated from the depositor input sheet is validated, all file links are checked against the actual deposit and it is made sure that all deposited files are referenced and described in the metadata. Additionally, some basic consistency checks on the information in the metadata files are performed and duplicates in the media files are detected. The generated reports with all open issues are then send back to the depositors to be rectified. Even though this process runs automatically, the LAC managers check the deposit report for plausibility and provide assistance and guidance during the whole process.

The metadata portal CLARIN Virtual Language Observatory[2] harvests the CMDI metadata of all CLARIN centres and displays the large amount of available resources through faceted browsing and search facilities. Both in the overview, i.e. when browsing or searching for relevant resources, and on the individual resource pages displaying further information on a specific resource, the user can report an issue or give feedback on metadata records or resources using a designated button connected via a form to the CLARIN-D Help Desk.

The CLARIN-D Help Desk manages support and feedback workflows for national centres and various international services, such as the CLARIN VLO. Depending on the type of feedback, help desk agents can thus both forward issues directly to the responsible CLARIN centre and, for issues with a wider impact, contact relevant institutions and bodies at the European level, such as the CLARIN Metadata Curation Taskforce, which is responsible for improving and harmonising metadata within the infrastructure.

Furthermore, the so-called discipline-specific working groups within the CLARIN-D project[3] are yet another communication channel, through which the various designated communities can provide more general input and feedback on data and metadata to ensure CLARIN-D centres provide relevant resources and resource descriptions.

References:
[1] <https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html>
[2] <https://vlo.clarin.eu>
[3] <https://www.clarin-d.net/en/disciplines>

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R12 Workflows

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The ingest and archiving processes are described in internal documentation[1] and are based on the OAIS reference model and described in the terminology established by the reference model.

Extensive documentation is provided for depositors. This includes depositor guidelines (https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/depositor-guidelines), submission guidelines (https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/submission-guidelines), and a format whitelist (https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/format-whitelist). The curation procedures are documented in the internal handbook and the current status of each dataset is tracked using a GitLab Issue Board.

There are two basic workflows: One for new datasets and one for legacy datasets.

When a new SIP is submitted via Sciebo, SOFS (the network storage provided by the RRZK) or via a physica storage medium, the archivist checks the deposit for obvious problems such as file formats, metadata format and content and the compliancy with the depositor agreement. Then the metadata contained in an Excel-Sheet is converted to the BLAM-metadata-format and enriched with normdata for languages[1], locations[2] and actors[3]. In an intial sanity check via a python script, the metadata is validated, file links are cross-checked against the deposited files and duplicates are detected. The generated reports with all open issues are then send back to the depositors to be rectified. This process is supported by the archivist.

Once the SIP passes the sanity check, the generation of the AIP is initiated. This includes the creation of the OCFL-folder structure[4], the creation of handles for all files and a final validation. The AIP is then copied to LAC-dev for manual inspection of its final appearance in the front-end. Should problems arise at this stage, a set of scripts performing automatic API-requests can be used to diagnose any open issues with the dataset. In a final step the dataset is uploaded to LAC-prod and a copy is moved to the long term storage system TSM.
For metadata dissemination, the repository uses a OAI PMH interface[3] (documentation[4]), which can be harvested by OLAC and the Virtual Language Observatory of CLARIN. Consumers can query the archive via these portals or via the web interface of the archive. Access to the DIPs is provided via the web interface of the archive. The DIPs can be delivered as a download or streamed in the browser.
Currently, there are three access levels implemented in the LAC:

- Open: Data can be access by anyone.
- Registration required: Consumers must be logged in with an XXXedu-roam-account or via the CLARIN identity provider (https://www.clarin.eu/content/clarin-identity-provider).
- Request required: Consumers must request access via email to the LAC-helpdesk who will contact the depositor for case by case decisions.

- Embargo: No Access is possible.

A more detailed description of the functional architecture and ingest pipelines of the DCH repository is available.[5]

References:
[1] https://redmine.uni-koeln.de/projects/ka3/wiki/Ingest_and_Archiving_Workflow (access protected, pdf version attached.)
[2] http://hdl.handle.net/
[3] https://api.ka3.uni-koeln.de/oai/lac
[4] https://ka3.uni-koeln.de/apidoc/oai
[5] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/archive-setup

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R13 Data discovery and identification

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The LAC provides search functionalities on its homepage[1]. Additionally, the archive provides a geographic representation of the archive’s content and supports map based browsing.

All CLARIN centres[2] provide their metadata in CMDI format. The Component Metadata Infrastructure (CMDI)[3] was initiated by CLARIN to provide a flexible framework for describing metadata based on components and concepts. Each metadata record is based on a profile that is registered in the CLARIN CMDI Component Registry[4]. Profiles can make use of components. Those building blocks are also registered in the CMDI Component Registry and describe specific aspects or properties of a resource. Elements of CMDI records link to concept definitions that are stored in external registries (like the CLARIN Concept Registry[5]). Since different communities use different names for the same concepts, linking CMDI elements to concepts enables communities to stick to their terminology while enabling users to find concepts independent of the naming.

A strict requirement for CLARIN centres is to make their metadata available through the established and well documented Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[6]. This standard enables harvesting of the metadata from the repository via http(s). The CLARIN Virtual Language Observatory (VLO)[7] harvests the metadata in CMDI format from all CLARIN centres via OAI-PMH. Metadata from all CLARIN centres (and other relevant archives and repositories) are browsable and searchable via the VLO website. CLARIN has defined a set of facets to narrow down the selection of resources in the VLO. These facets are again based on concept sets and allow access to potential heterogeneous metadata stocks. The search in the VLO combines a full text query with a selection of (multiple) values in facets.

Moreover, the LAC repository is also indexed by other registries. The LAC repository offers PIDs in form of a handle system and encourages to cite resources via their PIDs. To this end, the LAC has acquired a Handle prefix (11341) and operates a Handle server for persistent identifiers. The usage of PIDs is mandatory for resources and their CMDI metadata in CLARIN, thus all resources added to the repository can be referenced using PIDs. On the LAC-website all collections and bundles indicate the proper way to cite them.

References:
[1] <https://lac.uni-koeln.de/>
[2] <https://www.clarin.eu/content/overview-clarin-centres>
[3] <https://www.clarin.eu/content/component-metadata>
[4] <https://catalog.clarin.eu/ds/ComponentRegistry>
[5] <https://concepts.clarin.eu/ccr/browser/>
[6] <https://www.openarchives.org/pmh/>

[7] <https://vlo.clarin.eu>

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R14 Data reuse

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:

The guideline has been fully implemented in the repository - 4

Response:

The LAC repository closely follows the recommendations for standards and tools for compiling language corpora issued by Deutsche Forschungsgemeinschaft (DFG)[1]. The audio-visual collections archived with LAC are all available in digitized form and comply to a set of structural requirements and consist of whitelisted file formats. There are mechanisms to monitor potential format obsolescence and the repository continuously maintains a setup for migration.

The Data at the LAC is arranged in bundles. Bundles group closely related files such as an audio file and its transcription in an ELAN file. For the LAC, bundles are the main unit of archiving. Metadata are structured along the unit bundle and describe the bundles as well as the files a bundle contains of. This structure ensures, that related files are handled and recognized together and remain a unit.

The CMDI profile BLAM used by the LAC repository requires all essential metadata categories and enables the repository to deliver VLO, DataCite and OLAC compliant metadata. The OAI-PMH interface provides BLAM CMDI and Dublin Core metadata.

Only resources that comply with CLARIN guidelines and follow disciplinary and ethical norms are considered for deposit. The depositor is required to sign an agreement stating that these guidelines are met (see also R2 and R4).

Data sharing and reuse is promoted by providing access to the data (streaming and download) within the bounds of applicable licenses and free access to metadata (via the OAI-PMH protocol). The CLARIN service VLO[2] enables users to query combined catalogs containing metadata of all CLARIN repositories.

References:
[1] <http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf>, in German
[2] <http://www.clarin.eu/vlo/>

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Technology

R15 Technical infrastructure

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:

The guideline has been fully implemented in the repository - 4

Response:

The LAC repository complies with the OAIS reference model's tasks and functions. Moreover, the LAC's repository software is compliant with the Reference Model for an Open Archival Information System (OAIS). The LAC's repository software implements open interface standards such as the Open Data Protocol (ODATA), the International Image Interoperability Framework (IIIFS), OAI PMH, and the Security Assertion Markup Language (SAML) based Shibboleth system. The LAC with its partner RRZK developed the repository software within the joint BMBF funded KA³ project.

The infrastructure of the LAC is located at the RRZK. The base support is managed via the RRZK Infrastructure Level: The RRZK monitors its rate of hardware consumption on a frequent base. Due to regular applications the RRZK buys and installs new necessary hardware components (e.g. storage, computing-components, switches, etc.) dependant on how much, where and what is needed.

The infrastructure maintenance and development are supported by the projects the department performs. The RRZK provides the hardware and maintains virtual machines as well as the operating systems. The operation systems are kept up to date and always meet the current Long-Term-Support release of the used OS (Ubuntu).

The sources and the LAC's repository software itself are maintained utilizing a standard infrastructure. The source code is maintained in a Git repository located at the RRZK. A Redmine instance is connected to this Git repository and enables project employees to share information and document the software. The major part of the source code maintained in the Git repositories implements applications that rest on the grails framework. Grails is a framework based on the programming language Groovy, which is part of the Apache Software Foundation. It is a stable and reliable software development base. The Grails application framework has been used extensively for over ten years at RRZK department for service development and has been part of a long term expertise and sustainability strategy at the department. Locally developed Grails plugins are managed and stored in an Artifactory instance. A Docker repository manages and stores application server images.

The bandwidth sufficient available is adjusted in regular intervals. Currently the KA³ setup uses an infrastructure that is redundantly provisioned with a 1000 Mbit/s connection in contract with a local Internet provider.

As part of CLARIN-D we are committed to play an active role in the development of CLARIN's repository infrastructure. General plans for maintaining and further developing the infrastructure have been formulated as part of the project proposal. The central goal is to improve the usability of the research infrastructure for typical research tasks such as the retrieval of resources, the evaluation of data or the publication of results. To achieve this, modifications and extensions to a variety of infrastructure components in the repository and in the central infrastructure are necessary. Meetings of all centres to monitor advances in infrastructure development take place quarterly. Further important goals of infrastructure development are[1]:

- To ensure resilience, integrity, and availability of the sustainable repositories and the central infrastructure
- To integrate new resources and tools based on the needs of the user communities
- To allow for better interoperability of tools and resources in the infrastructure
- To enhance the central content search to be more useful in actual research tasks
- To optimize metadata of the resources provided and to enhance user experience in central metadata search Additional strategic infrastructure planning takes place on the European level in the coordinating committee of the technical centres of the CLARIN ERIC where CLARIN-D also participates.

References:
[1] <https://www.clarin.eu/content/clarin-technology-introduction>

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R16 Security

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

All data and metadata kept in the KA^3 repository of LAC are backed up several times a week to recent storage solutions (see R9). These storage solutions are co-located in geographically separated buildings in Cologne and are maintained by RRZK and its technical partners. Backup strategies and technologies are documented in an internal wiki.[1] This enables every authorized employee with administrative access to the storage infrastructure to fully recover a backup. In context of the continuous risk assessment undertaken by the LAC, the archive and its partners are evaluating additional sites to replicate the repositories data. At the current stage, legal and technical issues concerning remote location are being discussed. The repository is installed on a virtualized container cluster, which in turn runs in a scalable virtual machine subnet that is hosted on redundant hardware. The hardware and network are supervised via state-of-the-art software solutions and is additionally secured by ACLs maintained at the RRZK.

Access to each layers of the infrastructure is restricted to the respective department. So that for example only network administrators have access to network relevant administration, the storage department is in charge of handling storage related issues. The entire hardware stack is protected against unauthorized access and it is implemented within redundant power supplies, ventilation systems, to maintain temperature changes, fire warning and detection systems. The relevant departments and their contact information are available to the archive staff in the internal document.[2]

The infrastructure is permanently monitored via multiple tools, including Icinga, Ganglia, and OnCommand unified manager. The internal documentation provides information about the monitoring setup for the archive staff.[3] The university has an IT security officer as a central contact person, that is independent from the different IT services at the university. The business continuity of all central IT service of the university is ensured by the multiple uninterruptible backup power supplies (battery and diesel generator) for the technical infrastructure. Furthermore, the technical components in the server rooms of the RRZK are installed redundantly with power adapters in parallel and high-availability systems inside the server building. The LAC has a disaster recovery plan for scenarios affecting the repository or part of the infrastructure the repository depends on.[4] The plan provides a checklist for the archive staff. The archive staff are instructed to consult this document in case a malfunction requires recovering components or the complete system of the repository.

References:
[1] <https://redmine.uni-koeln.de/projects/ka3/wiki/Prozesse_im_Datenmanagement#Backup> (access protected, pdf attached)
[2] <https://redmine.uni-koeln.de/projects/ka3/wiki/Akteure_und_Rollen> (access protected, pdf attached)
[3] <https://redmine.uni-koeln.de/projects/ka3/wiki/System-Architektur#Monitoring> (access protected, pdf attached)
[4] <https://redmine.uni-koeln.de/projects/ka3/wiki/Disaster-Recovery> (access protected, pdf attached)

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Applicant Feedback

R17 Applicant Feedback

We welcome feedback on the CoreTrustSeal Requirements and the Certification procedure.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

-

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

-

Board comment: For recertification the board would expect improvements on the requirements that are now on compliance level 3.

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Thank you very much for addressing the reviewers’ comments, as well as for updating the text and many links. It is good to see more public, online, information supporting the application. Not only is this required for the certification, but it also allows your designated community, research funders and other stakeholders to convince themselves of the trustworthiness of the LAC repository.

Board comment: For recertification the board would expect improvements on the requirements that are now on compliance level 3.