Assessment Information

[CoreTrustSeal Requirements 2020–2024](#)

Repository:	CLARIND-UDS
Website:	http://fedora.clarin-d.uni-saarland.de/
Certification period:	June 13, 2024 - 13 June 2027
Requirements version:	CoreTrustSeal Requirements 2020-2022

This repository is owned by: **Universität des Saarlandes**

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background Information

Repository Type

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

Response:

- Domain or subject-based repository
- Institutional repository
- Research project repository

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Description of Repository

Provide a short overview of the repository.

Response:

The CLARIND-UdS centre (https://fedora.clarin-d.uni-saarland.de) is part of the Language Science and Technology department of Saarland University. It is not a legal entity of its own but an integral part of Saarland University.

In collaboration with other CLARIN centres, the CLARIND-UdS enables eHumanities by providing a service for hosting and processing language resources (notably corpora) for members of the research community.

CLARIND-UdS thus contributes to the overall CLARIN mission of reducing the fragmentation of language resources by assisting members of the research community in preparing language materials in such a way that easy discovery is ensured, interchange is facilitated and preservation is enabled by enriching such materials with meta-information, transforming them into sustainable formats and hosting them. CLARIND-UDS is a certified CLARIN-B centre (https://www.clarin.eu/sites/default/files/B-Centre%20Certificate%20-Universit%C3%A4t%20des%20Saarlandes.pdf, a renewal of the certificate requires the new CoreTrustSeal).

We have an explicit mission to archive language resources, especially multilingual corpora (parallel, comparable) and corpora including specific registers, both collected by associated researchers as well as researchers who are not affiliated with us.

CLARIND-UdS's director is a member of the NGO „Verein Geistes- und kulturwissenschaftliche Forschungsinfrastrukturen e.V." (https://forschungsinfrastrukturen.de/doku.php/start). This NGO represents the German CLARIN centres inside CLARIN through the German national coordinator (see https://www.clarin.eu/governance/national-coordinators-forum).

CLARIN (https://www.clarin.eu/) is an acronym for "Common Language Resources and Technology Infrastructure". It is a research infrastructure that was initiated from the vision that all digital language resources and tools from all over Europe and beyond are accessible through a single sign-on online environment for the support of researchers in the humanities and social sciences. The CLARIN infrastructure is fully operational in many countries, and a large number of participating centres are offering access services to data, tools and expertise.

In 2012, nine CLARIN member countries created CLARIN-ERIC (European Research Infrastructure Consortium), which is an international legal entity that governs and coordinates CLARIN activities, and it has grown to 22 full members and 2 observers since then. CLARIN-ERIC members are governments or intergovernmental organisations which pay an annual fee to support the development and maintenance of the CLARIN research infrastructure.

CLARIND-UdS is a part of the NFDI (Nationale Forschungsdateninfrastruktur "National Research Data Infrastructure") consortium (https://www.nfdi.de/consortia/?lang=en) Text+ (https://www.text-plus.org). The NFDI Text+ is one of finally 20 national research infrastructure consortia in Germany and it is funded via Deutsche Forschungsgemeinschaft (DFG). The NFDI Text+ will preserve text- and language-based research data in the long term and enable their broad use in science.

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Designated Community

Provide a clear definition of the Designated Community

Response:

The designated community of the CLARIND-UdS repository consists of linguists (specially corpus linguists), philologists, and researchers in digital humanities.

Reviews

Reviewer 1:

Comments:

Reviewer 2:

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

Response:

- 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

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Level of Curation - explanation

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

Response:

We don't do A—every content to be deposited is vetted by the repository staff.

Reviews

Reviewer 1:

Comments:

Reviewer 2:

CLARIND-UDS

Comments:

Insource/Outsource Partners

If applicable, please list them.

Response:

1) Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)

The repository makes use of a common CLARIN PID service (https://www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf) based on the Handle System (http://www.handle.net/) and in cooperation with the European Persistent Identifier Consortium (EPIC). CLARIN-D has a contractual relationship with GWDG concerning the provision of PID-services via EPIC API v2. The attached document lists the services which were stipulated. This outsource partner offers relevant functionality for Requirement 13 "The repository enables users to discover the data and refer to them in a persistent way through proper citation."

2) CLARIN-ERIC

CLARIND-UdS is a member of CLARIN's European Research Infrastructure Consortium (ERIC) through its membership in the NFDI Text+.
CLARIN-ERIC offers central services to its members and users, as stated here: https://www.clarin.eu/value-proposition The services are available to all centres in the member countries of the CLARIN-ERIC (https://www.clarin.eu/content/overview-clarin-centres).

The most important services of the ERIC cover the search functionality for the German CLARIN centres: Virtual Language Obervatory - VLO (https://vlo.clarin.eu): CLARIN's central metadata-based search engine, which contains metadata of all German CLARIN-centres.
Metadata harvester: The VLO is kept up to date using the metadata harvester run by the CLARIN- ERIC.
Federated Content Search - FCS (https://www.clarin.eu/contentsearch): Optionally, centres can provide the actual data of their resources for this central content search.
CMDI Component Registry (https://catalog.clarin.eu/ds/ComponentRegistry): CLARIN's registry for components and profiles according to ISO-24622-1.
In addition, CLARIN-ERIC offers several further services such as central registries, user statistics management and, as an official EUDAT community, access to advanced EUDAT services.

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Significant Changes

Summary of Significant Changes Since Last Application if applicable.

Response:

Since the last application, outsource partners and external relationships have changed, as listed in detail below. The scope and the mission of the repository remain untouched by those changes.
The funding of the CLARIN-D project has ended. All 8 CLARIN-D centres continue their operation and the mutual agreement to take over data from another centre in case of emergency is still valid. The CLARIN-D activities are now bundled in the NGO „Verein Geistes- und kulturwissenschaftliche Forschungsinfrastrukturen e.V." (see this recording from CLARIN Café
https://www.youtube.com/watch?v=h_E-cYXQNnE&list=PLlKmS5dTMgw1fEeU1ChH8_VldaR4ym8c1&index=6).
The CLARIND-UdS repository has joined the national research data structure (NFDI) Text+
https://www.text-plus.org/en/home/ and receives some funding from it. The funding scheme has changed from a pure project scheme with regular endings of individual projects to a funding subject to regular evaluation after 5 years.
The advisory boards of CLARIN-D do not exist any longer, their role has been overtaken by new advisory boards from NFDI Text+.
The membership of CLARIND-UdS in the European CLARIN project is not touched by these changes, the German contribution to CLARIN ERIC is paid by NFDI Text+.
The repository server was moved from HIZ Saarland (central IT for all universities of the state Saarland) to the IT group of the Department "Language Science and Technology" and is now maintained internally.

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Other Relevant Information

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

Response:

Within CLARIN, CLARIND-UdS 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.

The following requirements hold for CLARIN centres of type B, and are fulfilled by this resource center:

Centres need to offer useful services to the CLARIN community. Each centre needs to refer to CLARIN in a visible way on its website. Each centre needs to make explicit statements about its funding support state and its perspectives in this respect. Each centre needs to make explicit statements about CLARIN compliant resources and services available at the centre. Each centre needs to make clear statements about their policy of offering data and services and their treatment of IPR issues. The centre has to implement the GÉANT Data Protection Code of Conduct (DP-CoC) for each of its federated Service Providers. Centres need to have a proper and clearly specified repository system and participate in a quality assessment procedure as proposed by the CoreTrustSeal. 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 SAML2.0 and trust declarations. Centres need to offer component based metadata (CMDI) that make use of elements from accepted registries such as the CCR in accordance with the CLARIN agreements, i.e. metadata needs to be harvestable via OAI-PMH. Centres need to associate (handle) PIDs with their metadata records. These PIDs should be suitable for both human and machine interpretation, taking into account the HTTP-accept header. Individual files (e.g. a text, zip or sound file) can be referred to with either the PID of the describing metadata record in combination with a part identifier or with another PID. Centres can choose to participate in the Federated Content Search with their collections by providing an SRU/CQL Endpoint.

A short overview of all requirements for centres of type B is also given in the form of a checklist (https://www.clarin.eu/content/checklist-clarin-b-centres).

re3data record: https://www.re3data.org/repository/r3d100010384

DFG RIsources: https://risources.dfg.de/detail/RI_00435_de.html

CLARIN centre registry entry: https://centres.clarin.eu/centre/13

Reviews

Reviewer 1:

Comments:

Reviewer 2:

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:

We have an explicit mission [1] to archive and preserve language resources especially multilingual corpora (parallel, comparable) and corpora including specific registers, both collected by associated researchers as well as researchers who are not affiliated with us.

The mission is supported by the infrastructure of Saarland University and is integrated into NFDI Text+ and the international CLARIN infrastructures. As part of the CLARIN infrastructure, the repository is included in all promotional activities carried out at the national level by NDFI Text+ as well as the European level by CLARIN [2, 3, 4, 5, 6, 7].

For an overview of the mission and goals of the CLARIN research infrastructure, see the following publication by Erhard Hinrichs (former national coordinator of CLARIN-D) and Stephen Krauwer (former executive director of CLARIN-ERIC): Hinrichs, E.; Krauwer, S. (2014a): The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In: N. Calzolari et al. (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 1525–1531. ELRA, Reykjavík, Island. http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf

1. https://fedora.clarin-d.uni-saarland.de/index.en.html
2. https://www.clarin.eu/content/vision-and-strategy
3. https://office.clarin.eu/v/CE-2021-1830-ValueProposition.pdf
4. https://www.text-plus.org/events/korpora-in-text-kennenlernen-und-nachhaltig-nutzen/#more-2888
5. https://zenodo.org/record/6452002
6. https://www.text-plus.org/en/about-us/boards-2/
7. https://www.text-plus.org/en/about-us/participating-institutions/

English short summary for 4.: This is the announcement of a dissemination workshop from NFDI Text+ with a link to the program. The workshop is held in German language.

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

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:

All CMDI metadata are provided without access restrictions according to CLARIN recommendations. Part of the actual data is also provided without access restrictions, but a significant part of them is protected. The actual data have individual licenses and access restrictions. While the repository prefers data to be Open Access under a Creative Commons license, this cannot be granted for all archived resources. In one case it was necessary to arrange an in-house stay to comply with the license restrictions of a certain resource.

If applicable, the data consumer is made aware of usage restrictions for the data she/he has gotten access to. In most cases, the general usage restrictions are already described in the codes of conduct (see Requirement R4 for details), specific restrictions are specified by the depositor (if applicable). For some data, explicit statements need to be made by the data consumer about the usage of the data before he/she gets access. The depositor then decides on whether access is granted or not. In case of misuse, the only thing that can be practically done is to deny the user further access to the repository and to make the research community aware of the misuse.

Terms of Use: https://fedora.clarin-d.uni-saarland.de/termsofuse.en.html

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

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 CLARIND-UdS repository is mainly funded by Saarland University. In addition, it receives some funding from Deutsche Forschungsgemeinschaft (DFG) through NFDI Text+. The Text+ funding has a duration of 5 years and lasts until 30.09.2026. If evaluated positively, this funding scheme is going to be extended beyond that date.

In the case that the funding is completely ceased, Saarland University will provide access to the archived resources for a grace period.

In this time, a transfer of the resources to another archive can be negotiated. CLARIN centres commit to ensuring long-term availability, access and to preservation of datasets submitted to their repositories, as set out in their mission statements. CLARIN centres are set up as a distributed network, where each centre institution is a hub of the digital humanities and brings its own financial resources into the infrastructure ensuring continued availability. In case of a withdrawal of funding, the repository's content would be transferred to another CLARIN centre.

Links:

• [Memorandum of Understanding; Taking over research data of other CLARIN-D centres](#)

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

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:

Deposits are handled in a case-by-case approach. There are individual contracts and different licences for each resource we have archived. Access to the items is also handled case-by-case, ranging from open access over restricted access requiring a contract to restricted access onsite. The depositors themselves are responsible for compliance with any legal regulations in the area where the data is collected. In particular, all depositors are required to follow the The European Code of Conduct for Research Integrity [1].

Before ingest, the data are spot checked for proper anonymisation when disclosure risks are imminent. A description of our workflow for acceptance and curation is found in Requirement R12.

The repository will include resources provided by institutions related to NFDI Text+ and other institutions and/or organizations that belong to the extended Text+ community. The data in our repository contains sufficient information for others to assess the scientific and scholarly quality of the research data in compliance with disciplinary and ethical norms. We specifically rely on DFG ethical Codes of Conduct [2]. Our repository provides a quality assessment in that the data consumer can make some judgement about the level of trust or about the reputation of the depositor on the basis of the meta-information about the source institution/organization that is related to each resource. Our repository does not (and cannot) systematically verify whether the data received are collected according to these quality standards. We provide some guidance to depositors in terms of describing the full package of information that should be deposited to facilitate assessment at the repository's webpage [3,4]. The host of the repository, Saarland University, also has a guideline on "Principles of good scientific practice at Saarland University" [5] that applies to our repository.

1. ALLEA (ALL European Academies) European Science Foundation, The European Code of Conduct for Research Integrity, Revised Edition https://www.allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf

2. DFG, Guidelines for Safeguarding Good Research Practice
http://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice/index.html

3. https://fedora.clarin-d.uni-saarland.de/termsofuse.en.html

4. https://fedora.clarin-d.uni-saarland.de/depositors.en.html

5. https://www.uni-saarland.de/fileadmin/upload/verwaltung/ombudsperson/good_scientific_practice.pdf

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

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 CLARIND-UdS repository is run by the department of Language Science and Technology of Saarland University. It is represented by its director, Prof. Elke Teich (see evidence below).

The repository staff consists of scientists with solid knowledge and experience in the fields of corpus linguistics and metadata management. The staff dedicated to the repository consists of the project leader and one 50% FTE position for technology development, repository administration, and metadata curation. The staff members have access to training on data management, metadata curation, and long-term preservation offered by NFDI Text+ and CLARIN. The repository is supported by the internal IT group of the department.

In addition, there is another 70% FTE position financed by NFDI Text+ funds until the end of September 2026. There is also a budget for travel expenses to attend meetings, workshops, and conferences.

Our repository staff members have access to training on data management, metadata, long-term preservation and professional development (offered by Text+, NFDI e.V., and CLARIN). This includes regular meetings, mobility grants for sharing of expertise, conferences, meetings with their respective scientific communities as well as a centralized knowledge base. CLARIN has a wide field of expertise in its collaborative network of centres, which come from within their respective fields of digital humanities.

Links:

- Official University webpage of Prof. Elke Teich

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

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:

The CLARIND-UdS repository is run within the department of Language Science and Technology of Saarland University by scientists who are also members of the designated user community. The repository team is present at relevant conferences and also organises own workshops.

The repository, through its membership in NFDI Text+, is supported by external coordination committees [1]. The Scientific Coordination Committee (SCC) for task area Collections is the most relevant committee for CLARIND-UdS. The SCC receives the annual report from NFDI Text+ and gives feedback based on that report.

Also, through its membership in NFDI Text+, CLARIND-UdS is connected to a large number of Academic Societies (Fachgesellschaften)[2]. This guarantees close contact with the diverse communities and an understanding of their needs. The scientific associations play a central role in the governance structure of Text+, as they send representatives to the Text+ plenary and nominate candidates for election to the Coordination Committees.

[1] https://www.text-plus.org/en/about-us/coordination-committees-2/

[2] https://www.text-plus.org/en/about-us/academic-societies/

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

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:

The CLARIND-UdS repository does not allow for anonymous deposition of resources and it also does not provide any way for automated deposition by trusted users. Before a resource is accepted, the repository administration will be in contact with the depositor to assess the data, the metadata, and the legal issues. The repository administration offers guidance on how to produce good metadata and checks the formal correctness of them.

We consider all objects deposited in our repository as fixed and immutable. We do not host dynamically growing resources, and we do not allow deliberate changes in a deposited object. For protection against accidental changes in the saved datastreams, see Requirement 16. We create new digital objects for updates and keep the old versions in our repository.

The new version of a resource will contain a pointer to the older versions in its metadata.

The repository in principle makes the original deposited objects available in an unmodified way, if the objects were in one of the accepted file types and encodings. In the case of changes by the data producer, the repository creates a new digital object with a new persistent identifier. In the case that the repository has to change the data, e.g., because a file format becomes obsolete and superseded, the original data are kept.

The repository only accepts works from the original data producers, who are acknowledged as such by means of the "dc:creator" or "creator" elements, in Dublin Core or CMDI metadata respectively.

We use the Dublin Core field "dc:relation" in the metadata to maintain relations to other datasets, tools, or publications. The relations given there reflect the time when the resource was prepared and submitted and are contributed by the data producers.

We know the authors of our contributions from the scientific community and we are in contact with them during the ingest process. We do not formally check their identity.

Example

CMDI metadata record for the GRUG parallel treebank as delivered by the OAI provider

http://fedora.clarin-d.uni-saarland.de/oaiprovider/?verb=GetRecord&metadataPrefix=cmdi&identifier=oai:fedora.clarin-d.uni-saarland.de:clarind-uds:grug

An example of a versioned resource is the Royal Society Corpus that is currently archived in three versions,
Version 2.0 (http://hdl.handle.net/11858/00-246C-0000-0023-8D26-7), Version 4.0 (http://hdl.handle.net/21.11119/0000-0001-7E8B-6),
and Version 6.0 (http://hdl.handle.net/21.11119/0000-0004-8E37-F)

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

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 repository is in the implementation phase - 3

Response:

The data are provided in the formats chosen by the data producers from a list of
supported formats [1]. Metadata for each resource are always provided in both Dublin Core [2]
and CMDI [3] formats. The CLARIND-UDS repository requests good Dublin Core metadata from the
data provider and takes care of the conversion into CMDI.

The repository staff checks the metadata provided by the data depositor and asks for additional information
when the metadata are too sparse. The repository staff is also aware of the mapping of metadata to the
facets of the CLARIN Virtual Language Observatory (VLO) and suggests optimisations for better search results.

The repository did not experience any problems with non-preferred formats yet. We were always able to find
a consensus on an accepted format with our data providers.

The repository has not had the need to remove a resource until now. In this case, CLARIND-UdS will retain metadata of the resource telling that the data
in question is no longer available [see section "Withdrawal of Data" in 4].

At the date of the next submission the repository will have a Collection Development Policy that is coordinated with the other repositories in the NFDI
Text+. The plan describing this and other improvements for the next recertification can be found on the repository website (see evidence link "Schedule
for the further development of the repository").

References

1. https://fedora.clarin-d.uni-saarland.de/ressources/AcceptedFormats.en.pdf

2. https://www.dublincore.org

3. https://www.clarin.eu/content/component-metadata

4. https://fedora.clarin-d.uni-saarland.de/depositors.en.html

Links:

* [Schedule for the further development of the repository](#)

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

Comments:

R9 Documented storage procedures

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The deposited data are stored in-house on a running virtual machine. They are secured against

* physical loss
* change of the data
* unauthorized access

We provide a public documentation of the data preservation policy [1]. The archived data are stored on a virtual machine run on a distributed hardware environment in two different buildings. Regular backups are taken and the integrity of the data is tested using checksums. For more information see requirement R16.

Metadata in CMDI format is created and optimised in close collaboration with the depositor to ensure an adequate description of the data; for more information see our answer on requirement R14.

[1] https://fedora.clarin-d.uni-saarland.de/preservation.en.html

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

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 CLARIND-UDS repository performs data curation on levels B (basic curation) and C (enhanced curation). Basic curation is applied before the ingest of every resource, enhanced curation is done on a case-by-case basis when a data format is becoming obsolete. The depositors grant the right to convert dataformats to the repository [1, 2].

Measures are taken to enhance the chance of future interpretability of the data. The number of accepted file formats is limited, to make 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 to ensure future interpretability.

Before ingest, we do the following checks:

* Full validation of the metadata against their respective schemes (oai_dc, CMDI)
* Check that provided XML data are well-formed (at least)
* Spot-check of character encoding

All archived resources are available online, the access permissions are defined by the data producers/depositors themselves.

The crisis management plan relies on the technical solution described in Requirement R9 of these guidelines.

The repository will provide a plan for future migrations of obsolescent data formats with the next re-certification. The plan describing this and other improvements for the next recertification can be found on the repository website (see evidence link "Schedule for the further development of the repository").

Links:

CLARIND-UDS

- [Preservation Policy](#)
- [Depositors' Agreement](#)
- [Schedule for the further development of the repository](#)

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Reviewer 2:

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:

The CLARIND-UdS repository 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. The CLARIN Curation Dashboard [1] carries out on-demand and automated quality checks for CMDI metadata profiles, records and collections. It also verifies the availability of resources referenced from harvested metadata.

Researchers interested in including their resources into the CLARIND-UdS repository are invited to develop a data management plan [2] in coordination with the staff of the CLARIN centre, which is offered as a service already in early states of their projects. Otherwise, actual data currently is only accepted when data depositors can be trusted (see R8: Appraisal). Metadata may be submitted by the data depositor or will be created in coordination with the CLARIND-UdS centre - in any case, it will be thoroughly checked and discussed by experienced centre staff. We encourage the data depositors to cite related work, most notably a publication about the provided resource, in the metadata they provide using the dc:relation element from Dublin Core.

The metadata portal CLARIN Virtual Language Observatory [3] 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 Helpdesk.

[1] https://curation.clarin.eu/collection/UdS_CLARIN_D_Repository.html

[2] https://www.clarin-d.net/de/aufbereiten/datenmanagementplan-entwickeln

[3] https://vlo.clarin.eu

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

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 CLARIND-UDS repository complies with the OAIS reference model's tasks and functions [1]. Besides, the repository is powered by Fedora Commons software, which is
compliant with the Reference Model for an Open Archival Information System
(OAIS) due to its ability to ingest and disseminate Submission Information
Packages (SIPS) and Dissemination Information Packages (DIPS) in standard
container formats.
We provide a structure diagram of the repository [2]

The ingestion workflow is defined by the archive management tool Fedora Commons, such as that no resource can be archived without metadata and that the resource has to conform to certain file formats and encodings. The responsibilities of the depositor are

* to decide what kind of material is being archived,
* to assure that the material follows the technical criteria required by the repository,
* to decide who may access the material,
* to protect the privacy of any subjects appearing in the recordings or texts.

There is an internal documentation on the preparation of resources and the corresponding metadata. For the time being there is no need to make these publicly available as we do not intend to implement an automatic ingestion process.

The data depositors grant the repository the licence to convert the submitted data to other formats for the purpose of data curation [3].

The data consumer has direct access to the archived objects via the web, provided that access requirements have been met. In the case of protected resources, special arrangements can be made on individual basis, including an in-house stay to access a resource that cannot be given out due to copyright restrictions.

For details on the search interfaces for dissemination, see our answer to requirement R13.

References

1. Reference Model for an Open Archival Information System (OAIS),
Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2,
June 2012 https://public.ccsds.org/pubs/650x0m2.pdf

2. https://fedora.clarin-d.uni-saarland.de/struktur.en.html

3. https://fedora.clarin-d.uni-saarland.de/depositors.en.html

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

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 data are provided in the formats chosen by the data producers from a list of supported formats, see section 2 of this guidelines for the full list of supported formats. Metadata for each resource are always provided in both Dublin Core and CMDI (Component MetaData Initiative [1]) formats.

Search facilities over metadata are available at our repository [2]; but a much more user-friendly search over our metadata is provided by the Virtual Language Observatory (VLO) [3]. 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. Since we cooperate with the VLO within the framework of CLARIN, we don't plan any improvement of our local search interface.

Harvesting of metadata is implemented via OAI-PMH. Our OAI provider [4] offers such metadata for OAI-PMH [5] harvesting in two formats: Dublin Core and CMDI.

The repository itself does not offer a persistent identifier service on its own but makes use of PID service offered by GWDG [6] based on the handle system [7]. We register handles from the handle service as persistent and resolvable identifiers for our resources.

Furthermore, the repository provides a section for data users [8], where links to search interfaces, data user agreement and citation good practices are provided.

References

1. https://www.clarin.eu/content/component-metadata

2. http://fedora.clarin-d.uni-saarland.de/fedora/objects

3. CLARIND-UdS browsing facet at VLO:
https://vlo.clarin.eu/?fq=collection:Universit%C3%A4t+des+Saarlandes+CLARIN-D-Zentrum,+Saarbr%C3%BCcken

4. UdS OAI-PMH provider: http://fedora.clarin-d.uni-saarland.de/oaiprovider/?verb=Identify

5. https://www.openarchives.org/pmh/

6. CLARIN's PID short guide: https://www.clarin.eu/sites/default/files/pid-CLARIN-ShortGuide.pdf

7. Handle system: http://www.handle.net

8. http://fedora.clarin-d.uni-saarland.de/users.en.html

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

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 repository is in the implementation phase - 3

Response:

The data producer is required to produce metadata accomplishing the data formats provided by the repository. CLARIND-UdS requests metadata according to the Dublin Core [1] standard. In the process of ingest, these metadata are semi-automatically converted to CMDI [2] by means of an XSLT template [3] and enriched with additional information, e.g., persistent identifiers (PIDs). CLARIND-UdS provides a list of accepted formats [4, see also 6], which includes common multimedia-document formats as well as formats for binaries. For other file formats, we provide advice for conversion.

We deliberately use the freedom of Dublin Core metadata to leave fields blank or to provide them several times. We insist on metadata on the following topics

* author of the resource
* title of the resource
* description of the resource
* rights

We also ask for data on the other basic Dublin Core fields, whenever available.

We manually check the metadata provided by the data producer and ask for additions or clarifications when the metadata are not sufficient. We also offer help and guidance on metadata writing.

We currently use exclusively the CMDI profile clarin.eu:cr1:p_1288172614026 (OLAC-DcmiTerms) [5]

Besides the metadata that describe the resources we hold there are also in-resource metadata describing the texts inside a corpus or the tokens inside a text [7]. Those metadata are explained individually on the resource landing pages, for an example see [8]. With this documentation the data are understandable and the resource is reusable in research.

We have not yet had the demand for upgrading a resource for a new data format – in fact, sometimes we were asked to provide simpler data formats for some specific resources, and we offered that downgraded versions on demand.

For the next re-certification, this section shall include a discussion of community use cases and how the repository deals with the community's demand. The plan describing this and other improvements for the next recertification can be found on the repository website (see evidence link "Schedule for the further development of the repository").

[1] https://dublincore.org/

[2] https://www.clarin.eu/content/component-metadata

[3] https://www.clarin.eu/faq/how-can-i-convert-my-dc-or-olac-records-cmdi

[4] https://fedora.clarin-d.uni-saarland.de/ressources/AcceptedFormats.en.pdf

[5] https://catalog.clarin.eu/ds/ComponentRegistry/#/?itemId=clarin.eu:cr1:p_1288172614026

[6] https://www.clarin.eu/content/standards-and-formats

[7] http://dx.doi.org/10.32714/ricl.09.01.02

[8] https://fedora.clarin-d.uni-saarland.de/rsc_v6/annotation.html

Links:

* [Schedule for the further development of the repository](#)

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

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:

CLARIND-UdS runs on a LAMPJ (Linux, Apache (httpd and tomcat), MySQL, Python, Java) software stack. We maintain a software inventory in an internal Wiki and keep up with the necessary software updates. The repository complies with the OAIS reference model's tasks and functions [1]. Besides, the repository is powered by Fedora Commons software, which is compliant with the Reference Model for an Open Archival Information System (OAIS) due to its ability to ingest and disseminate Submission Information Packages (SIPS) and Dissemination Information Packages (DIPS) in standard container formats. The Archival Information Package (AIP) consists of the metadata and their history stored in Fedora Commons, the checksums for the datastreams stored in the configuration file of our own checksum checker, and the list of registered handles stored in our internal wiki.

The hardware environment and the backup procedure are described in the answer to R16. Saarland University has a high bandwith connection provided by DFN (Deutsches Forschungsnetz [2]).

The data consumer has direct access to the archived objects via the web, provided that access requirements have been met. We provide a structure diagram of the repository [3]. We are committed to play an active role in the development of the repository infrastructure. 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. An example for such a modification is an XSLT stylesheet that gives a human readable view of the CMDI metadata of a resource in the CLARIND-UdS repository [4].

Further important goals of infrastructure development are (https://www.clarin.eu/content/clarin-technology-introduction):

* 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 Text+ also participates.

References

1. Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2, June 2012 http://public.ccsds.org/publications/archive/650x0m2.pdf

2. https://www.dfn.de/

3. http://fedora.clarin-d.uni-saarland.de/struktur.en.html

4. http://fedora.clarin-d.uni-saarland.de/ressources/olac2cmdi2html.xsl

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

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:

The deposited data are stored on a running virtual machine. They are secured against

* physical loss
* change of the data
* unauthorized access

Protection against loss of data

The repository is run in the virtual machine environment provided by the internal IT. The virtual machine environment is installed redundantly on two hardware clusters located in two different buildings on the Saarbrücken Campus. The internal IT takes four snapshots a day and keeps them for 95 days. Some snapshots are kept for a period of up to two years, in addition every six month backups on hard disk drives are taken and preserved in a fire-resistant vault [1].

Protection against change of data

The repository monitors the integrity of all archived files using two kinds of checksums (MD5 and SHA1). The checksums are controlled on a daily basis, and in case of a checksum mismatch a backup copy of the resource will be restored.

Protection against unauthorized access

The repository is guarded against unauthorised access by a firewall that exposes only http and https to the outside. It runs on a dedicated virtual machine with restricted access from the the internal network.

CLARIND-UdS is certified as a CLARIN-B centre [2,3]. There are strict criteria to become a CLARIN B-Centre: it should be based on a stable technical and institutional foundation. The Assessment Committee checks these requirements during an assessment procedure, while the technical coordination among the centres takes place in the Centre Committee. One security criterion is the existence of an SSL certificate for all servers involved that provides a full trust chain. The technical risks (like hardware failure) are handled by the internal IT. Recovery is done (and it has been actually done in the past) by the repository staff.

1. https://fedora.clarin-d.uni-saarland.de/preservation.en.html
2. https://www.clarin.eu/content/certified-centres
3. https://www.clarin.eu/node/3767

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

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:

We added a schedule for the further development of the repository. According to this schedule, we will deal with R8 in the second quarter of 2024, with R10 in third quarter of 2024, and finally with R14 in the last quarter of 2024. The link to the schedule was also added to requirements R8, R10 and R14.

Links:

* [Schedule for the further development of the repository](#)

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Board comments:
* Future recertification - Compliance levels (general): For recertification in 3 years the board expects that the applicant demonstrates significant progress on requirements previously assessed at a compliance level below the maximum obtainable level. That is, applicants are expected to progress to "fully implemented".

Reviewer comments: Thanks for the improvements. This re-certification has three requirements (R8, R10, R14) remaining at compliance level 3. As mentioned in the review fields this is acceptable with the condition that the repository makes explicit what extra steps will be taken to get to level 4 in the next re-certification.

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Board comments:
* Future recertification - Compliance levels (general): For recertification in 3 years the board expects that the applicant demonstrates significant progress on requirements previously assessed at a compliance level below the maximum obtainable level. That is, applicants are expected to progress to "fully implemented".

Reviewer comments: This renewal is quite meticulous; information and links are updated or edited accordingly.

There are mentions of new evidences, such as Clarin Curation Dashboard and in-resource metadata, which are excellent.

In this re-certification, a few requirements are on compliance level 3. Clarind-UDS has provided a scheduled plan which depicts the milestones that it has to achieve in 2024.