Assessment Information

CoreTrustSeal Requirements 2017–2019

Repository:	CLARIN-D Resource Center Leipzig
Website:	http://clarin.informatik.uni-leipzig.de/repo/
Certification Date:	19 February 2019

This repository is owned by: **University of Leipzig**

CLARIN-D Resource Center Leipzig

Notes Before Completing the Application

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

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository, Research project repository

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

CLARIN-D Resource Center Leipzig (http://clarin.informatik.uni-leipzig.de/repo/) is part of a network of research infrastructure centers located throughout Europe. The aim of these centers is to provide language data, tools and services in an integrated, interoperable and scalable infrastructure to researchers in the humanities and social sciences.

CLARIN-D Resource Center Leipzig is a domain and subject-based repository
focussing on written text corpora, general lexical resources and resources for lesser resourced languages.
It offers unique data collections of past and ongoing research projects, mostly corpora of the Leipzig Corpora Collection and related tools, making it also a research project repository.
Additionally, the repository allows individuals and organizations to deposit collections that are considered worthwhile preserving for future generations and research projects.

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. CLARIN-ERIC members are governments or intergovernmental organisations which pay an annual fee to support the development and maintenance of the CLARIN research infrastructure.

Germany is one of the founding members of CLARIN-ERIC and contributes to it via CLARIN-D (https://www.clarin-d.net/en/). CLARIN-D is an acronym for “Common Language Resources and Technology Infrastructure Deutschland”.

CLARIN-D Resource Centre Leipzig (http://clarin.informatik.uni-leipzig.de/repo/) is one of currently eight German CLARIN-D Resource and Service Centres which form a web and centres-based research infrastructure for the social sciences and humanities. The center is part of the CLARIN-D consortium funded by the German Federal Ministry for

Education and Research. The aim of CLARIN-D and its service centres is to provide linguistic 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 building on the achievements of the preparatory phase of the European CLARIN initiative as well as CLARIN-D's Germany-specific predecessor project D-SPIN. These previous projects have developed research standards to be met by the CLARIN services centres, technical standards and solutions for key functions, a set of requirements which participants have to provide, as well as plans for the sustainable provision of tools and data and their long-term archiving.
This repository offers resources such as a set of corpora of the Leipzig Corpora Collection (http://corpora.uni-leipzig.de/), based on newspaper, Wikipedia and Web text. Furthermore several REST-based webservices are provided for a variety of different NLP-relevant tasks.

Within CLARIN, this resource centre 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.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository’s Designated Community.

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 this field, CLARIN-D Resource Center Leipzig, is mainly providing resources and tools for scholars of lexicography, general corpus linguists or typologists working with resources of a large number of languages.

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).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

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

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

For all deposited resources at least basic curation is provided.

Collections may be deposited by third parties deciding to retain full ownership of the data. In this case they also assume responsibility for data curation activities, such as future migrations to new data formats. We perform basic checks on the metadata and data (e.g. completeness, validity and checksums) and may request additional information if deemed relevant for maintenance, discoverability and usability of the collection and guide and support the depositor in these tasks.

If feasible, depositors are further supported by our staff. This mainly involves support for conversion of the actual data into appropriate formats. The process of choosing the final format takes place in close collaboration with the depositor and members of the user community from CLARIN's working groups in the respective fields of the humanities.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Outsource Partners. If applicable, please list them.

1) Gesellschaft fu■r Wissenschaftliche Datenverarbeitung mbH Go■ttingen (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). The usage of PIDs is mandatory for resources in CLARIN thus all resources added to the repository may be referenced using PIDs. CLARIN-D has a contractual relationship with GWDG concerning the provision of PID-services via EPIC API v2. The following document lists the services which are stipulated:
http://www.clarin-d.de/mwiki/images/0/0b/GWDG_PID.pdf

2) CLARIN-D
The repository in one of currently eight resource and service centres of CLARIN-D. As part of the CLARIN-D consortium, the repository has signed the "Konsortialvertrag" - Cooperation Agreement - which states the rights and obligations of all CLARIN-D centres. A condensed version of this contract (in German only) is available at:
https://www.clarin-d.net/de/ueber/zentren/zusammenarbeit

CLARIN-D offers several services to it's member institutions, among them the following:
- CLARIN-D HelpDesk (https://support.clarin-d.de/mail/): A central system for user support, which allows for the distribution of user questions and feedback to qualified personnel at the centres.
- CLARIN-D website (https://clarin-d.de/en/): A starting point for researchers to find information on CLARIN-D and to access CLARIN-D services.
- CLARIN-D wiki (https://www.clarin-d.de/mwiki/index.php/Hauptseite): A central platform for CLARIN-D-related staff.
- CLARIN central monitoring (https://monitoring.clarin.eu/): A monitoring service offered to all CLARIN-ERIC members and maintained by the resource centre Leipzig.

3) CLARIN-ERIC
CLARIN-D is a member of CLARIN'S European Research Infrastructure Consortium (ERIC). CLARIN-ERIC offers central services to it's 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.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

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.

An overview of all requirements for centres of type B is also given in the form of a checklist (https://office.clarin.eu/v/CE-2013-0095-B-centre-checklist-v6.pdf).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

ORGANIZATIONAL INFRASTRUCTURE

I. Mission/Scope

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

The mission of the CLARIN-D Resource Centre Leipzig is to ensure the availability and preservation of resources, to preserve knowledge gained in research, to aid the transfer of knowledge into new contexts, and to integrate new methods and resources into university curricula. It serves as the repository of a CLARIN-D resource center of type B (http://www.clarin.eu/files/centres-CLARIN-ShortGuide.pdf – approved by CLARIN coordinator's office).

This mission is supported by the infrastructure of the University of Leipzig and by the integration of the repository into the national and international CLARIN infrastructures. As part of the CLARIN-D infrastructure, it shares the CLARIN-D mission to provide linguistic data, tools and services in an integrated, interoperable and scalable infrastructure for the Humanities and Social Sciences (https://www.clarin-d.net/en/about), and is committed to play an active role in the development of

CLARIN's repository infrastructure.

For a more general overview of the mission and goals of the CLARIN research infrastructure, see the following publication by Erhard Hinrichs (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

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

II. Licenses

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

Currently only data that is available for free (open data, completely publicly available) and states this in a license or comes with “compatible” licenses will be added to the repository. Currently, mainly data created by the institution which runs the repository is present. External depositors have to sign a depositor agreement. These contracts contain statements on

(1) the involved parties

(2) licenses and copyright

(3) rights and responsibilities of the depositor and the repository

(4) the content to be deposited

(5) removal of content and access conditions

(6) availability to third parties

(7) provisions relating to use by third parties

(8) death of the Depositor

(9) liability

(10) term and termination of the Agreement

This document is available on the repository website (http://clarin.informatik.uni-leipzig.de/repo/).

Depositors need to sign an agreement stating that they own all necessary rights required to deposit the data and that during the creation of the resource the data producer respected IPR (Intellectual Property Rights) and privacy issues. Data depositors are themselves responsible for compliance with any national or international legal regulations. Since no data with disclosure risk will be added to the repository, depositors also have to state that the deposited resource does not contain any data with disclosure risk. The repository staff maintains a checklist of cases in which resources containing data with disclosure risk have previously been rejected or modified (and if so, how they were modified) in order to be compliant to the repository regulations. This list is intended to help in cases in which the depositors are unsure about the status of their resource regarding disclosure risk.

In case a violation of conditions is observed, the original data provider is contacted. In case the violator can be identified, further access by this person/institution will be prevented if technically possible. If feasible, the violator and the violator’s home institution will be contacted on the issue personally.

The system does not allow the integration of data into the repository without providing an appropriate license. These license conditions are available to the users via CMDI metadata. 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.

Despite the repository only accepting open data, a document on access permissions can be found on the repository's website:

http://clarin.informatik.uni-leipzig.de/repo/files/access_permissions_v2.pdf

This document also includes information on how non compliance with the accass permissions is handled.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

III. Continuity of access

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

In the depositor agreement it is stated:

“

The Repository shall ensure, to the best of its ability and resources, that the deposited
Content is archived in a sustainable manner and remains legible and accessible.
The Repository shall, as far as possible, preserve Content unchanged in its original
digital format, taking account of current technology and the costs of implementation.

“

The repository is not a legal entity on its own. It is run by the Natural Language Processing Group of the University of Leipzig which is an institution governed by public law. In addition it is part of the CLARIN-D research infrastructure consortium.
CLARIN-D and therefore the repository is funded by Bundesministerium für Bildung und Forschung (BMBF) with a project based funding for terms of four years. Additionally, the repository is funded by the University of Leipzig in conjunction with the Leipzig Corpora Collection.

All 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 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. In this case, the funding by the University of Leipzig can at least ensure the intermediate-term maintenance of the infrastructure of the Centre. 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 (https://www.clarin-d.net/en/about/centres/mou-taking-other-centre-s-data). The legal aspects of the process of relocating data to another institution is addressed by templates of license agreements provided in CLARIN.

The repository is currently in the process to develop its future strategy allowing to guarantee preservation periods to the data depositors. Hence discussions with BMBF, state of saxony and University of Leipzig are ongoing. Additional funding to allow for a closer collaboration with DARIAH-DE research infrastructure (https://de.dariah.eu/en/startseite) is available from March 2019 until March 2021. A proposal for a participation in a national research data infrastructure (Nationale Forschungsdateninfrastruktur - NFDI) for the humanities is currently created.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IV. Confidentiality/Ethics

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

Depositors need to sign an agreement (which might additionally be tailored to fit the needs of a certain resource or depositor) stating that they own all necessary rights required to deposit the data and that during the creation of the resource the data producer respected IPR (Intellectual Property Rights) and privacy issues. In particular, data must be anonymized when applicable. Problematic aspects of each resource are determined in close coordination with each data depositor. Since the number of externally provided resources in the repository is low, an elaborate procedure is possible in these cases.

Data depositors are themselves responsible for compliance with any national or international legal regulations. Since no data with disclosure risk will be added to the repository, depositors also have to state that the deposited resource does not contain any data with disclosure risk. The repository staff maintains a checklist of cases in which resources containing data with disclosure risk have previously been rejected or modified (and if so, how they were modified) in order to be compliant to the repository regulations. This list is intended to help in cases in which the depositors are unsure about the status of their resource regarding disclosure risk.

In case a violation of conditions is observed, the original data provider is contacted. In case the violator can be identified, further access by this person/institution will be prevented if technically possible.
Guidelines are also provided for users of the repository:
http://clarin.informatik.uni-leipzig.de/repo/files/access_permissions_v2.pdf

Users are requested to ensure ethical use of all resources. In case misuse is identified by the staff of the repository or the staff is informed by external researchers or CLARIN-D's working groups in the humanities, appropriate actions are taken. In such a case of noncompliance with disciplinary and ethical norms first of all the violator will be contacted to ensure the misuse is stopped immidiately.

No data with disclosure risk will be submitted to the repository, but still compliance with disciplinary and ethical norms must be ensured. In case of doubt, the depositor is contacted on the issue. Additionally, experts of CLARIN-D's discipline-specific working groups in the humanities will be contacted.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

V. Organizational infrastructure

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

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Reviewer 2

Comments:
3 – The repository is in the implementation phase

Response:

The repository is hosted by the Natural Language Processing Group of the University of Leipzig. Hosted within the Computer Science Department it is ensured that its staff has sufficient knowledge in this central field of operating and developing a repository. Additionally students of the NLP-group are requested to minor in linguistics or related fields to allow for sufficient knowledge when interacting with scholars of our target group, the humanities, or when working on different types of language resources or tools.

In addition our repository is part of CLARIN-D, a research infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences. CLARIN-D also offers information on a wide range of topics, including teaching material, help on data management plans and other, discipline-specific support.

By being part of the CLARIN-D consortium the repository gains access to funding for running and further developing a sustainable repository and resource centre to support these goals. Besides staff resources this includes a budget for attending different national and international meetings such as conferences or workshops. For its employees CLARIN offers training on data management, metadata, long-term preservation and professional development (offered by CLARIN-D and CLARIN-ERIC). This includes regular developer meetings, mobility grants for sharing of expertise, conferences, meetings with their respective scientific communities (called discipline-specific working groups) as well as a centralized knowledge base (user guide, wiki, bugtracker and mailing lists). CLARIN has a wide field of expertise in its collaborative network of centres, which come from within their respective fields of digital humanities.

Currently CLARIN-D is funded by the Bundesministerium für Bildung und Forschung (BMBF). The current project phase has a runtime of 4 years and is funded until 30.09.2020. The university of Leipzig and the Natural Language Processing Group contribute additional funding. Altogether four full time positions are attributed to running and maintaining the repository and to the integration of tools and language resources. The tasks of the staff members are stated in the "Arbeitsplan" of CLARIN-D. An overview is available at https://www.clarin-d.net/en/about/centres/division-of-labour-of-clarin-d-centres.

As an alternative to project based funding, CLARIN-D currently pursuits a permanent continuation of funding. Discussions and negotiations on this topic are ongoing. A proposal for a national research data infrastructure (Nationale Forschungsdateninfrastruktur - NFDI) for the humanities is planned.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VI. Expert guidance

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

The repository, through its membership in 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: Reinhard Altenhöner, Christiane Fellbaum, Björn Granström, John Nerbonne, Heike Renner-Westermann, and Achim Streit.

The joint Technical Advisory Board (TAB) of CLARIN-D and DARIAH-DE is a committee 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:
Jan Haji■ (Prague Institute of Formal and Applied Linguistics, CLARIN Center), Margareta Hellström (ICOS Carbon Portal staff member), Peter Leinen (German national library, head of information technology), Karlheinz Mörth (Austrian Acacemy of Sciences), Wolfgang Nagel (Technical University of Dresden, Head of the centre of information services and high performance computing), and Sabine Roller (University of Siegen, head of the centre of information and media technologies).

How does the repository communicate with its Designated Community for feedback?
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 other related 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 (https://www.clarin-d.net/en/disciplines/). For communication, mailing lists and wiki contents are maintained.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:

Accept

DIGITAL OBJECT MANAGEMENT

VII. Data integrity and authenticity

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

We documented the process of depositing and ingesting resources as a BPMN 2.0 process model (see http://clarin.informatik.uni-leipzig.de/repo/files/Process_Ingest.pdf). As can be seen, there is close cooperation between the depositor and repository staff in case questions or problems arise. The depositor has to submit a defined set of documents along with the data:

- the signed Resource Deposition Request Form (RDRF)
- a checksum for the submitted data
- metadata corresponding to the submitted data

After reviewing the RDRF document regarding completeness, we might ask the depositor to complete the document within a defined period of time. After all RDRF issues have been resolved, we will check the RDRF along with the data regarding several aspects, including the identity of the depositor, the origin of the data and licensing terms. In some cases we will reject the submission instantly, e.g. if the type of data is inappropriate – in other cases we will ask the depositor for further clarification or modification of the submission. If no more issues remain, the submitted metadata are validated and the checksums are verified. In case metadata validation is unsuccessful, fixed metadata will be requested. If any submitted checksum does not match the checksum we calculated for the data that was submitted, we will ask the depositor to resubmit the data.

After successfully passing all checks, the data is imported into our Version Control System (VCS). The submitted checksums will be verified against the imported data afterwards. In case of a mismatch, a staff member will fix any errors that might have changed the data and run the import again.

After successfully importing the data and verifying checksums, a unique persistent identifier (PID) will be created and updated in the metadata if it does not already contain such an identifier. This PID will be used to unambiguously identify the resource.
The following steps will ensure the accessibility of the submitted resources. While the backup process of VCS resources is in progress, the metadata will be imported into a Fedora Commons repository system. For performance reasons, the Fedora repository is supported by a cache which has to be updated after new data has been imported. Once the update process is done we will verify the integrity of the imported metadata by matching it against the metadata contained in the VCS. In case of any inconsistencies we will remove the faulty metadata, fix remaining problems and repeat the procedure. At the end of the process, the depositor will receive a notification.

We implemented the use of checksums in order to monitor the integrity of digital objects stored in the repository. We perform these kinds of checks when:
(1) new data is added to the repository (data retrieved from the repository needs to be identical to the data that was ingested)
(2) periodically (e.g. every time a backup is created) in order to ensure that no data was changed unintentionally

For this, Apache Subversion (SVN) has been set up as part of our repository. Within this SVN the data of the repository is already stored as a backup mechanism. These backups are created once the integrity of the data in the repository was ensured after ingestion. In addition a checksum of the original data is created and stored. A mechanism for regular comparison of the state of the resources in the SVN and the repository to checksums created upon insertion into the version control system was implemented.

These mechanisms are closely linked to the storage procedures described in R9 such as the backup process (http://clarin.informatik.uni-leipzig.de/repo/files/Process_Backup.pdf).
Access to data and metadata is provided via webservice interfaces. The availability of these webservices is monitored via Icinga (https://www.icinga.org/) probes. Some of these probes are run in local installations at the center while others are operated by CLARIN-D (http://clarin-d.net/images/ap3/ap3-005-monitoring.pdf). The frequency of checks depends on the type of service that is monitored.

Multiple versions of data are valid. CLARIN propagates the idea of reproducible research. Thus updates/new versions of existing data is handled like any other resource with the exception of setting and storing a reference to the previous version. Access to metadata and data of all versions is provided at the same time and is handled in the same way:
(1) access is provided via OAI-PMH/webservices
(2) and a unique PID is assigned.

However, updates of metadata for existing resources are possible without considering the result to be a new version. 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 parsed for syntactic correctness and manually evaluated for completeness and soundness.

In case data that is present in the repository “changes”, this data is considered to be a new version of the existing data. Thus data producers need to provide the same type and scale of information (metadata, documentation) that was provided for the previous version (at least in case changes occurred).

The metadata provided for each resource to be added to the repository needs to contain basic information on the data depositor (e.g. name of the institution, contact address) and the provided data (e.g. name, date or version, description of the resource itself and of the data format being used, obligatory links to papers). Adding further information (e.g. change logs) is encouraged but not enforced. In case multiple version of a resource are present in the repository, at least references to previous/newer versions needs to be present in the metadata.

Data and metadata are essential and mandatory parts of the digital objects that represent a resource in the repository. This can be considered to be an implicit link between data and metadata. In CMDI metadata is explicitly linked to data and additional metadata via the ResourceProxy-section (https://www.clarin.eu/faq/3462) in a CMDI file.

Currently we do not intend to compare essential properties of different versions of the same file/resource. Keeping track of changes that occurred in between different versions of the same file/resource will be up to the data producers. In order to improve the usability we will encourage but not enforce data producers to provide change-logs in case new versions of already existing data are ingested into the repository.

Currently there is no explicit check of the identity of depositors. So far, all depositors were met in person or were previously known to the staff of the repository from the context of CLARIN. Once this changes an explicit procedure for the check of depositor identities and “ownership” of the ingested data needs to be specified. External deposits will only be accepted after a due diligence process involving a check of the identity of the depositor and a clarification of all legal issues.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VIII. Appraisal

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

The repository is closely related to the Projekt Deutscher Wortschatz / Leipzig Corpora Collection (LCC) and to a large share is submitted to making LCC-resources available to research communities. The staff of the repository is also active in the development of the LCC, as according to the “Arbeitsplan” of CLARIN-D. Therefore, just like the LCC, the repository has a focus on written text corpora, reference corpora, general lexical resources and resources for lesser resourced languages. Preferably resources from these fields or of high scientific value for the respective communities are integrated into the repository. Primarily, resources for integration are chosen based on relevance for their respective field. This is also stated on the repository's website (https://clarin.informatik.uni-leipzig.de/repo/files/resource_depositor_guidelines_v4.pdf).

For corpora of the Leipzig Corpora Collection, which regularly are added to the repository, a selection process for choosing priorities is in place based on the following aspects:

- Is textual data in the language already available in the repository?
- How many speakers does the language have (data available from ethnologue: https://www.ethnologue.com/)?
- Is the quality of the data high enough or can it be assessed at all?

CLARIN-D strives to guide depositors to a fitting resource centre. On it's central website the project offers a guide for finding such a resource centre based on the properties of the data which are to be deposited: https://www.clarin-d.net/en/preparation/find-a-clarin-centre.
On this website the Resource Centre Leipzig is present as well and is connected to topics such as “Lexical data, web services and special reference-corpora, public data“.

The handling of requests to deposit data that does not fall within the (CLARIN) mission will be decided on a case by case

basis. Data that supports the CLARIN mission and that is of high relevance to the respective communities will be prioritized.
Quality control checks to ensure the completeness and understandability of data are based on the requirements of the repository for resources to be deposited.
The minimal requirements for data/tools to be deposited in the repository are:
(1) The data/tool is provided in a standardized format or with an exhaustive documentation of the proprietary format
(2) Metadata is available in CMDI
(3) contact information on the data depositor / data producer is present in the metadata
(4) a statement on the legal status of the resource is available

The data that is put into the repository is checked for compliance with internal and CLARIN guidelines concerning scientific and scholarly quality. Only data that:
(1) is the result of research projects,
(2) comes with exhaustive metadata,
(3) which's data structure is described by a sophisticated documentation (PDF/A),
(4) comes with information on how the data was originally created,
(5) was reviewed by a third party
will be added to the repository. The data itself, the metadata and additional documentation is an obligatory part of each repository entry.

Currently these guidelines are not in a fixed state and subject to minor changes. They are available on the repository website (http://clarin.informatik.uni-leipzig.de/repo/).

The data stored in the repository is mostly well known and documented content created by our own institution as the result of a long running research project. For external resources we check the complience with the requirements and guidelines in close collaboration with the data providers. Integration will only take place in case all requirements are met. If possible, the depositor will be supported in the process of resolving any issues, but the responsibility remains with the depositor.

Metadata for all CLARIN repositories has to be provided in the CMDI format. There is exhaustive documentation (http://www.clarin.eu/cmdi) available on how to create CMDI compliant metadata profiles and instances. Additionally a set of tools is provided that allow data producers to easily create new or adapt existing metadata to the CMDI standard. Resources must be accompanied with valid CMDI metadata in order to be considered for deposit. Metadata is checked for compliance according to CMDI standards in the following way:
(1) Check if XML metadata is well-formed and valid.
(2) Check if the used CMDI components and profiles are stored in the Component Registry (http://catalog.clarin.eu/ds/ComponentRegistry/) (public, PID present).
(3) Check if the data categories used in those components/profiles are present in the CLARIN Concept Registry (https://concepts.clarin.eu/ccr/browser/).
(4) Check if the provided CMDI files contain enough and consistent information (e.g. consistent specification of the data producer's “name”) according to the needs of the VLO (http://www.clarin.eu/vlo/).

The granularity of CMDI metadata is up to the (meta)data producer. The repository itself is able to handle a high granularity of metadata. The creation of metadata files (instances) is supported via the XML Editor ARBIL (https://tla.mpi.nl/tools/tla-tools/arbil/) that comes with CMDI support. Metadata elements need to be compliant to the standards set in CMDI. Since CMDI is a component based approach which allows (meta)data producers to create custom tailored metadata profiles there is no limit to the usage of established standards etc. In order to be visible and useable in the CLARIN infrastructure CMDI metadata added to the repository needs to contain a minimum set of attributes (linked to data categories stored in the CLARIN Concept Registry) which is enforced by the quality checks described above. The usage of metadata elements that are accepted by a research community is encouraged and technically supported via re-use of existing metadata components (created in close collaboration with the respective communities in CLARIN's working groups in the humanities), but is not enforced.

This information is part of the resource depositor guide which is available on the repository website (http://clarin.informatik.uni-leipzig.de/repo/).

In case the the metadata provided are insufficient for long-term preservation, the data will not be added to the repository. In this case, the aim of our repository is to support the depositor in providing the missing metadata and including it in the final dataset.
It is recommended to use formats listed in the CLARIN standard recommendations (http://www.clarin.eu/recommendations). In addition relevant standards and formats in the context of CLARIN are listed (http://www.clarin.eu/content/standards-and-formats). These lists were generated in close collaboration with the respective communities of different fields of the humanities. Manual checks are performed before data is added to the repository. Usage of standardized formats is encouraged but not enforced.
In case no recommended/well known and documented format is used, an exhaustive documentation on the syntax and semantic of the data (e.g. database dumps: names of tables and columns; specifications and examples on the contents of each column; examples on how to retrieve different types of data) will have to be provided by the data producer. This documentation (English, PDF) will be stored on the repository along with the data and metadata and is provided to everyone who wishes to download/access the resource. The repository maintainers keep track of all formats already used by the depositors and commit themselves to work on updates of the CLARIN standard recommendation if new formats gain in popularity.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IX. Documented storage procedures

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

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Reviewer 2

Comments:
3 – The repository is in the implementation phase

Response:

Relevant procedures for storage are documented in an internal Mediawiki instance.
In the preservation policy basic information on data storage is available.
http://clarin.informatik.uni-leipzig.de/repo/files/ULei_preservation_policy_v2.pdf

The repository runs servers in the computing centre of the University of Leipzig. The necessary storage is provided by redundant storage systems. The machines are housed in a modern server rooms, which provide air conditioning and (limited) uninterrupted power supplies. Access to the server rooms is limited to authorized staff. Maintenance of the systems is performed by the repository's system administrator. On these servers, virtual servers are hosted in the server virtualization management platform Proxmox. Access to the virtual servers is restricted by a firewall. The storage hardware and hardware for virtual machines is replaced at regular intervals and based on information extracted from monitoring data.

Data is stored on RAID systems and backups are created on a regular basis (every time the content of the repository changes, since ingests are done by the repository maintainers). These backups are held on separate hardware, located in separate server rooms in a separate fire safety zone of the university of Leipzig. We are currently discussing the topic of additionally keeping backups at external computing centres which are partners of CLARIN-D, but results are still open and legal, technical, and financial issues still need to be settled. Together with the documented workflows the existing backups enable full recovery of the repository.

The integrity of the data elements is ensured via SVN. Additionally for backups of full virtual servers checksum tests are in place.
Deterioration of storage media is monitored via Icinga probes which do a regular check of the used hardware (e.g.

S.M.A.R.T. - Self-Monitoring, Analysis and Reporting Technology - data) and report drastic changes or imminent failures. In case of failures/problems/... the administrators of the repository are notified and will take appropriate actions.

Backups of Virtual Machines are created regularly using the Proxmox backup mechanisms. This process is documented in a BPMN 2.0 diagram (see http://clarin.informatik.uni-leipzig.de/repo/files/Process_Backup.pdf) which is also referenced in the ingestion process (see http://clarin.informatik.uni-leipzig.de/repo/files/Process_Ingest.pdf). During the backup process, a VMA file and its checksum are created for the repository machine (see https://pve.proxmox.com/wiki/VMA). Once this backup has been stored on a separate storage, checksums are verified. Once the problems causing the data alteration have been resolved, the backup is created again.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

X. Preservation plan

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

External depositors have to sign a depositor agreement. These contracts contain statements on

(1) the involved parties

(2) licenses and copyright

(3) rights and responsibilities of the depositor and the repository
(4) the content to be deposited
(5) removal of content and access conditions
(6) availability to third parties
(7) provisions relating to use by third parties
(8) death of the Depositor
(9) liability
(10) term and termination of the Agreement

A preliminary version (which will be subject to change based on future experience with depositors) is available on the repository website (http://clarin.informatik.uni-leipzig.de/repo/).

The data provider retains all intellectual property rights to their data. The depositor must grant distribution rights to the repository. Access is provided by the repository and distribution rights are specified in the written agreement. Enforcing licenses by data users in the case of misuse is conducted by the property rights owner.

Crisis management concerning the availability of the digital objects is addressed on a technical level (described in R9). Since a PID system is used in CLARIN, moving resources from one CLARIN resource center to another one is possible without affecting the validity of references (e.g. PID of a resources used in a paper). Our setup consists of virtual machines which may be moved to other CLARIN partners . In case virtual machines are moved internally (inside the CLARIN-D center in Leipzig) this will be possible without severe impact to user experience (live migration is supported). In case the machines need to be moved to other CLARIN partners a limited downtime will occur. Legal aspects of the process of relocating data to another institution is addressed by templates of license agreements provided in CLARIN.

By encouraging data depositors to use standardized formats (UTF-8, documented XML formats, ...) we try to minimize the cases in which obsolescence of file formats will occur in the near future. By enforcing a detailed and exhaustive documentation in case proprietary / “custom” formats are used we ensure that exhaustive documentation is available under all circumstances. Thus it will, at least, be possible to specify and implement data converters.

Long term data usability is ensured by the following measures:
(1) We make sure that all data formats, also proprietary ones, are well documented.
(2) We enforce provision of information on authorship of the data and encourage adding references to scientific papers describing the data and usage scenarios.
(3) Access to data and metadata is provided via widely used open source software stacks (MySQL, Tomcat, Fedora Repository) that are installed on virtual machines. This maximizes the probability of long term support (updates, security fixes) for the tools being used and improves the ability to run installations of these software stacks independent from the underlying hardware/operating system/... .

For further information please refer to the preservation policy provided on the repository website (http://clarin.informatik.uni-leipzig.de/repo/).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XI. Data quality

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

The repository Leipzig has established specialized workflows to ensure the quality of both data and metadata during creation, ingestion and dissemination (see R7, R9 and R12). Those workflows are executed and monitored by adequate staff based on designated roles in the repository's internal organizational structure.

Staff members have long-term experience with their assigned workflows, hosted resource types and the scientific environment. This includes long-term experience in the creation, management and publication of lexical resources and text corpora for the targeted communities. Most staff members of the repository have both a computer science background and are active researchers in the field of corpus linguistic and text mining. They have therefore a thorough understanding of the needs and requirements of the research communities. Most staff members are also members of relevant taskforces and working groups of the CLARIN project, including those about metadata standardization and metadata quality.

Quality of resources is evaluated before any ingestion activities take place. These include semi-automatic evaluation procedures based on resourcetype-specific checklists and language-statistics based detection of anomalies and outliers in the text data (if applicable, see Thomas Eckart, Uwe Quasthoff und Dirk Goldhahn: Language Statistics-Based Quality Assurance for Large Corpora. In: Proceedings of Asia Pacific Corpus Linguistics Conference 2012, Auckland, New Zealand, 2012). Part of the ingestion workflow is also a manual inspection of the resource by personnel having experience with the respective resource type. For resources provided by external depositors, the depositors are encouraged to provide the repository with proof of the data quality, relevance of the resource and known shortcomings (like in form of peer-reviewed publications).

To ensure quality of metadata, the repository uses strict schema validation for all provided metadata records (for both resources of the project Deutscher Wortschatz and deposited resources). In regular intervals (typically bi-annual) the metadata schemata are manually evaluated for their fitness and adequacy. If a demand for upgrades and revisions are identified, both schemata and metadata records are improved (see R14). The repository has already identified shortcomings in the past and has functioning workflows for upgrading and dissemination procedures.
As a means of external control and supervision, the quality of metadata records are investigated during the CLARIN centre assessment every three years. As an automatic tool to ensure and improve metadata quality, the CLARIN project provides the CLARIN Curation Module that continuously monitors provided metadata of all associated repositories - currently more than 60 - and prepares an evaluation using a variety of quality measures (like validness of records, accessibility of contained URLs, adequacy for presentation in search engines, etc). The Leipzig repository typically ranks among the Top-5 repositories based on a combined score of all evaluated features (https://curate.acdh.oeaw.ac.at/#!Collections).
Further information about the curation module can be found at:
https://office.clarin.eu/v/CE-2016-0742-CLARINPLUS-D2_1.pdf.

To provide end users with sufficient information about the quality of data the repository uses two approaches in parallel.
The repository provides easy-to-use Web interfaces to get access to relevant statistics, data and data samples, like
- statistical information about various features of the corpora provided via the CLS portal (https://cls.corpora.uni-leipzig.de/en)
- easy-to-use Web interfaces to access sample data for manual inspection
- providing references to peer-reviewed publications about the data and their quality (if available)

Furthermore, the repository makes uses of the general CLARIN infrastructure that supports evaluation and fast feedback:
- for all resources detailed metadata records are provided via a standard interface (OAI-PMH). Those records are accessible for end users in the faceted metadata search engine Virtual Language Observatory (VLO, https://vlo.clarin.eu).
- support of end users via the feedback and reporting function of the VLO, which are forwarded to the responsible CLARIN centre
- support of end users using the CLARIN-D Help Desk, where help desk agents forward questions or remarks directly to the responsible CLARIN centre
- support and feature requests by the discipline-specific working groups of the CLARIN-D project

(https://www.clarin-d.net/en/disciplines) 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

For all three communication channels, dedicated and qualified personnel is assigned.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XII. Workflows

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

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Reviewer 2

Comments:
3 – The repository is in the implementation phase

Response:

Currently, necessary workflows that define how to integrate/archive data provided by external data providers are established. Based on experiences with data integration from external sources the creation of such workflows is in progress. A prelimenary version of the process of depositing and ingesting resources as a BPMN 2.0 process model (see http://clarin.informatik.uni-leipzig.de/repo/files/Process_Ingest.pdf) has been created and might be further modified. The general backup process is documented in a BPMN 2.0 diagram (see http://clarin.informatik.uni-leipzig.de/repo/files/Process_Backup.pdf).

The establishment of these workflows currently goes hand in hand with upgrading the backend of the repository from the most recent version of Fedora Commons 3 to the most recent version of Fedora 4. In this process typical workflows for integrating internal resources and experiences with integrating external resources are utilized.

Final workflows are expected to include at least the following technical validation and automated curation while ingesting CMDI metadata and the underlying data, including:

1. Validation against CMDI Schemas before the ingest.
2. Integrity check for all referenced data.
3. Generation of an URL for CMDI records and data, and registration of the URL in the handle system (http://hdl.handle.net/).
4. Validation based on the validation procedures of the underlying Fedora-Commons backend.
5. Validation of CMDI Records delivered by the OAI Provider.

We have created initial documentation on how to archive types of internal resources that are regularly added to the repository. Based on this work this documentation will be extended in order to address similar kinds and other types of resources we expect to be added by external depositors. Once all currently open questions mentioned below are part of the documentation, these documents will be available on the repository website (http://clarin.informatik.uni-leipzig.de/repo/).

As already mentioned in R8 (Appraisal), the repository has a focus on written text corpora, reference corpora, general lexical resources and resources for lesser resourced languages. Preferably resources from these fields or of high scientific value for the respective communities are integrated into the repository. Primarily, resources for integration are chosen based on relevance for their respective field.

For corpora of the Leipzig Corpora Collection, which regularly are added to the repository, a selection process for choosing priorities are in place based on the following aspects:

- Is the language already available in the repository?
- How many speakers does the language have (data available from ethnologue: https://www.ethnologue.com/)?
- Is the quality of the data high enough or can it be assessed at all?

CLARIN-D strives to guide depositors to a fitting resource centre. On it's central website the project offers a guide for finding such a resource centre based on the properties of the data which are to be deposited.
https://www.clarin-d.net/en/preparation/find-a-clarin-centre

While some questions still need to be answered (e.g. up to which scale are we able to handle big data of external depositors) some outlines are already clear from a CLARIN perspective:

(1) Only open data, data that is freely available or data that is free to be used for research and teaching will be added to the repository.
(2) There needs to be a plausible usage scenario and a user community for the deposited data/tools.
(3) If possible, access to the data has to be provided via webservices on a level of granularity that matches these

use-cases.

(4) Metadata in CMDI needs to be present.

The handling of requests to deposit data that does not fall within the (CLARIN) mission will be decided on a case by case basis. Data that supports the CLARIN mission will be prioritized.

Reviewer Entry

Reviewer 1

Comments:
For future certifications, the workflow description should preferably be available in a public document.

Reviewer 2

Comments:
Accept

XIII. Data discovery and identification

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

All CLARIN centres (https://www.clarin.eu/content/overview-clarin-centres) provide their metadata in the CMDI format. The Component MetaData Infrastructure (CMDI) (https://www.clarin.eu/content/component-metadata) 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 (https://catalog.clarin.eu/ds/ComponentRegistry). 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, https://openskos.meertens.knaw.nl/ccr/browser/). 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, and therefore for the Resource Centre Leipzig as well, is to make metadata for all resources available through the established and well documented Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (https://www.openarchives.org/pmh/). This standard enables harvesting of the metadata from the repository via http(s).

Main search facilities are currently not provided by the repository itself. Instead, services of the CLARIN-ERIC are utilized. The provision of harvesting services for metadata and the provision of central metadata and data search facilities is stated in the value proposition: https://www.clarin.eu/value-proposition

Metadata:

The CLARIN Virtual Language Observatory (VLO) (https://vlo.clarin.eu) of the CLARIN-ERIC harvests the metadata in CMDI format from all CLARIN centres via OAI-PMH. Metadata from 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.

Data:

For a subset of resources of the CLARIN-infrastructure a “deep search” within the actual data is supported by the means of the CLARIN Federated Content Search (http://www.clarin.eu/fcs) interface. CLARIN-D Resource Center Leipzig also offers this kind of access for some of its resources, especially for the corpora of the Leipzig Corpora Collection.
In addition, the repository offers it's own portal for lexical searches in the datasets of the Leipzig Corpora Collection. It is available at: http://clarinportal.informatik.uni-leipzig.de/en

PIDs:

The repository uses the 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). The usage of PIDs is mandatory for resources in CLARIN thus all resources added to the repository may be referenced using PIDs.
CLARIN has a contractual relationship with GWDG concerning the provision of PID-services via EPIC API v2 as mentioned in R0 on repository context.
For each resource of the Leipzig Corpora Collection, which is made available in the CLARIN context, PIDs are available: on corpus level as in the following link: http://hdl.handle.net/11022/0000-0000-23CC-3

on sentence level as in the following link:
http://clarinws.informatik.uni-leipzig.de:8080/cmdi2/cmdi/11022/0000-0000-23CC-3?type=dataprovider&id=39
PIDs for both levels are easily accessible on the resource's page in the VLO.

Citation:

Necessary information for citing resources of the repository - such as PID, resource name, and responsible organization - can be found on the respective page of the VLO (as well as in the metadata):
https://vlo.clarin.eu/record?5&docId=hdl_58_11022_47_0000-0000-23CC-3
https://vlo.clarin.eu/record?5&docId=hdl_58_11022_47_0000-0000-23CC-3&tab=technical

We are currently in the process of adding information on recommended data citations to the corpus portal (http://clarinportal.informatik.uni-leipzig.de/en). In addition, discussions with CLARIN-ERIC have been initiated on integrating recommended data citations into the VLO.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIV. Data reuse

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

Metadata for all CLARIN repositories has to be provided in the CMDI format. There is exhaustive documentation (http://www.clarin.eu/cmdi) available on how to create CMDI compliant metadata profiles and instances. Additionally a set of tools is provided that allow data producers to easily create new or adapt existing metadata to the CMDI standard.

Resources must be accompanied with valid CMDI metadata in order to be considered for deposit. Metadata is checked for compliance according to CMDI standards in the following way:
1. Check if XML metadata is well-formed and valid.
2. Check if the used CMDI components and profiles are stored in the Component Registry (http://catalog.clarin.eu/ds/ComponentRegistry/) (public, PID present).
3. Check if the data categories used in those components/profiles are present in the CLARIN Concept Registry (https://concepts.clarin.eu/ccr/browser/).
4. Check if the provided CMDI files contain enough and consistent information (e.g. consistent specification of the data producer's "name") according to the needs of the VLO (http://www.clarin.eu/vlo/).

In order to be visible and useable in the CLARIN infrastructure CMDI metadata added to the repository needs to contain a minimum set of attributes (linked to data categories stored in the CLARIN Concept Registry) which is enforced by the quality checks described above. The usage of metadata elements that are accepted by a research community is encouraged and technically supported via re-use of existing metadata components, but is not enforced. A document containing a summary of metadata fields which are mandatory or seen as desirable by the Leipzig repository is provided (http://clarin.informatik.uni-leipzig.de/repo/files/deposit_requirements_metadata_v1.pdf).

CMDI profiles which are currently in use at the repository are documented at the Component Registry. Currently, the following profiles/schemata are in use:
LCC_CorpusProfile (clarin.eu:cr1:p_1381926654510)
LCC_DataProviderProfile (clarin.eu:cr1:p_1381926654510)
WebLichtWebService (clarin.eu:cr1:p_1320657629644)
OLAC-DcmiTerms(clarin.eu:cr1:p_1288172614026)
Metadata schemata and the corresponding metadata records are regularly updated and extended when the requirement for additional information arises. For metadata schemata that are not in the responsibility of the repository Leipzig (like WebLichtWebService or OLAC-DcmiTerms), their further development is promoted in the designated CLARIN taskforces and working groups.

It is recommended to use formats listed in the CLARIN standard recommendations (http://www.clarin.eu/recommendations). In addition relevant standards and formats in the context of CLARIN are listed (http://www.clarin.eu/content/standards-and-formats). These lists were generated in close collaboration with the respective communities of different fields of the humanities. Manual checks are performed by CLARIN members before data is added to the repository. Usage of standardized formats is encouraged but not enforced.

In case no recommended/well known and documented format is used, an exhaustive documentation on the syntax and semantic of the data will have to be provided by the data producer. This documentation (English, PDF) will be stored on the repository along with the data and metadata and is provided to everyone who wishes to download/access the resource.

The repository maintainers of CLARIN-D keep track of all formats already used by the depositors and commit themselves to work on updates of the CLARIN standard recommendation if new formats gain in popularity. This is done in conjunction with CLARIN's standards committee (https://www.clarin.eu/governance/standards-committee). For this purpose, the close collaboration of CLARIN's resource centres and the scholars from the humanities in working groups or during dissemination events is very helpful. In case migration of existing resources seems necessary, the depositor of a data set will be contacted.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

TECHNOLOGY

XV. Technical infrastructure

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

For metadata we rely on the group of standards around CMDI (ISO-CD 24622-1).
The repository complies with the OAIS reference model's tasks and functions. Moreover, the repository is based one the Fedora Commons software, which is compliant with the Reference Model for an Open Archival Information System (OAIS).
For this repository the OAIS
(1) Submission Information Package (SIP) consists of:
(1.1) the (binary) data to be stored in the repository
(1.2) metadata in CMDI that further describes the resources
(1.3) a documentation or specification on the formats beeing used (links to documentation in case of standardized formats or exhaustive documentation on the format in case a proprietary one is use)
(2) Archival Information Package (AIP) consists of:
(2.1) all information/data that is part of the SIP
(2.2) a persistent identifier for the resource (usually obtained by the repository)
(2.3) an overall checksum

Metadata is available in CMDI via OAI-PMH. The CMDI file of a resource contains links to documents stored in the repository, interfaces - usually webservices in CLARIN – or webapplications that facilitate usage of the resource. The CMDI file tied together with these resources can be seen as a representation of a Dissemination Information Package (DIP).

Within CLARIN, used standards are discussed and reviewed on a regular basis. The standards committee is reponsible for their adoption. https://www.clarin.eu/governance/standards-committee

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 (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 CLARIN-D also participates.

System documentation is available to all staff in our internal Mediawiki instance. This documentation is currently reviewed and extended for more completeness of necessary information. It includes, among other aspects:

-hardware details of all servers (including age of components)

-relevant software and services (including operating system) of servers and virtual machines

-information on backups

-set up guides

Additionally, the resource centre is operating a Git repository, which contains all relevant software projects of the repository together with thorough documentation.

The repository is using well-supported operating systems and additional core infrastructural
software which is appropriate to the services it provides to its Designated Community.

As operating systems instances of Debian Linux are used: https://www.debian.org/

Proxmox VE, a complete open-source platform for enterprise virtualization, is utilized for virtualizing our machines: https://www.proxmox.com/en/

We use Fedora Commons Repository 3 as our main repository solution: https://github.com/fcrepo3

We utilize MySQL as database solution: https://www.mysql.com/

The services of the repository are running on three main physical servers:

A) Production system with a Proxmox Ve host: all service/websites/webapps run on dedicated virtual systems

B) Backup Production system with a Proxmox Ve host: Allows for quicker recovery of the repository in case of a failure of the main host

C) Data Backup System: Physical server with over 100TB for offside Backups of: VM-Snapshots and database-dumps

All physical servers are equipped with RAID systems.

On the Productions system (A) there are 5 dedicated VMs for different tasks. For each VM once a week full snapshots are created (last four snapshots will be stored) and stored on the data backup system:

1) clarin: provides CLARIN Repository Website and portal for direct access to the corpora

2) clarinoai-fedora: Fedora Repository

3) clarinws: Provides Webservices for the Websites and Webapps

4) clarinarchive: MySQL-Server supply data to clarinws

5) clarindevel-current: GIT-repository for repository related code

All Vms are running on a Debian 9.X Operation System with several protection mechanisms like restricted IP-Ranges for SSH/MySQL-Access, fail2ban and Firewall/ufw.

In order to provide maximum uptime, VM Snapshots can easily be restored to one of the Proxmox Ve Nodes.

All physical and virtual servers are monitored by an icinga instance, which performs several checks like APT, LOAD, MEM, HDD and PING.
This icinga instance is running independent from the additional central monitoring of the CLARIN-ERIC and allows the repository to monitor more performance and system-metrics. The central instance (CLARIN-ERIC) is mainly used for uptime checking and availability checks of standardized interfaces. In addition the local icinga2 instance offers the information necessary to guarantee the proper functionality of all local services and hardware.
To guarantee that all Software-Components are up-to-date the local monitoring checks (using APT) if new packages are available in the linux-software repository. Once the admin is notified by icinga, the respective systems get updated.
With the checks LOAD and MEM we monitor the workload of all VMs, so that we can react to overload. This could e.g. be caused by automatic querying of services, which could then be stopped by banning of respective IPs.
To prevent system-fails because of missing HDD-space the HDD-check is necessary to getting a Warning before a HDD capacity is exhausted.
On a regular basis (twice a year) monitoring reports based on icinga data exports are analyzed by the technical staff of the repository. Problematic aspects are then discussed and appropriate actions are taken. Especially statistics based on LOAD, MEM and HDD are of interest. They can reveal upcoming shortcomings of the technical infrastructure and, if necessary, lead to the replacement or expansion of hardware.
Remark: No real-time to near real-time data streams are provided by the repository.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XVI. Security

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

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Reviewer 2

Comments:
3 – The repository is in the implementation phase

Response:

The repository runs servers in the computing centre of the University of Leipzig. The necessary storage is provided by redundant storage systems. The machines are housed in a modern server rooms, which provide air conditioning and (limited) uninterrupted power supplies. Access to the server rooms is limited to authorized staff. Maintenance of the systems is performed by the repository's system administrator. On these servers, virtual servers are hosted in the server virtualization management platform Proxmox. Access to the virtual servers is restricted by a firewall. The storage hardware and hardware for virtual machines is replaced at regular intervals.

Data is stored on RAID systems and backups are created on a regular basis (e.g. weekly and every time the content of the repository changes, since ingests are done by the repository maintainers). These backups are held on separate hardware, located in a separate fire safety zone within the University of Leipzig.

The integrity of the data elements is ensured via SVN. Additionally for backups of full virtual servers checksum tests are in place.
Deterioration of storage media is monitored via Icinga probes which do a regular check of the used hardware (e.g. S.M.A.R.T. - Self-Monitoring, Analysis and Reporting Technology - data) and report drastic changes or imminent failures. In case of failures or problems the administrators of the repository are notified and will take appropriate actions.
All hosting services of the repository are running on virtual machines whose images are regularly backed up. In addition to the primary VM host an additional backup host has been set up. The backup server is also situated in the University of Leipzig, but in a different fire safety zone, to reduce risks of full data loss. In case of a failure of the main host swift recovery will be achieved by running the backed up VMs on the backup host. All services of the repository are monitored via a local instance of Icinga. In addition a central CLARIN-Icinga is in place to monitor the central services of each repository such as OAI-PMH-endpoints. Therefore, the staff of the repository will be informed immediately via Email in case of outages and can react accordingly.
Additionally, the repository is the responsible centre for monitoring activities in CLARIN-D. Hence, the repository staff is creating reports on the reliability of its own services and the services of other CLARIN-centres for internal CLARIN-D-meetings and to keep track of the status of the infrastructure.

Currently, no published disaster plan or business continuity plan exist. Documented workflows for typical recovery activities exist and are available on the repositories website. Together with the existing backups they enable full recovery of the repository.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

APPLICANT FEEDBACK

Comments/feedback

These requirements are not seen as final, and we value your input to improve the core certification procedure. To this end, please leave any comments you wish to make on both the quality of the Catalogue and its relevance to your organization, as well as any other related thoughts.

Response:

Thank you for taking the time to assess our repository. Your work is greatly appreciated.

Please inform us if further explanations or documentation is necessary.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept