Assessment Information

CoreTrustSeal Requirements 2020–2023

Repository:	CLARIN Center Leipzig
Website:	https://repo.data.saw-leipzig.de
Certification period:	22 August 2023 - 21 August 2026
Requirements version:	CoreTrustSeal Requirements 2020-2022
This repository is owned by:	**Saxon Academy of Sciences and Humanities in Leipzig**

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background Information

Repository Type

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

Compliance level:

Not Applicable - 0

Response:

- Domain or subject-based repository
- Research project repository

Links:

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Description of Repository

Provide a short overview of the repository.

Compliance level:

Not Applicable - 0

Response:

The repository is a member of the European CLARIN research infrastructure and the research infrastructure consortium Text+ which is a member of the German initiative to establish a national research data infrastructure (Nationale Forschungsdateninfrastruktur – NFDI).

The mission of both, CLARIN and Text+, is to create an infrastructure that makes language resources and language technology readily available and usable to scholars of all disciplines, in particular the humanities and social sciences. In this field, the CLARIN repository 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. This includes written text corpora, reference corpora, general lexical resources and linguistic resources. Underlying time periods depend on the respective datasets, but are often based on modern contexts such as Web text. A special focus of this repository lies on so-called "under-resourced" or "under-represented" languages.

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.

The Text+ infrastructure is focused on language and text data and will initially concentrate on digital collections, lexical resources and editions. These are of high relevance for all language- and text-based disciplines, especially for linguistics, literary studies, philosophy, classical philology, anthropology, non-European cultures and languages, as well as language- and text-based research in the social, economic, political and historical sciences. The repository is one of Text+'s data and competence centres, focusing on the Text+ data domain "Lexical resources" and taskforce "Infrastructure/Operations".

Links:

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Designated Community

Provide a clear definition of the Designated Community

Compliance level:

Not Applicable - 0

Response:

The CLARIN and Text+ mission is to create a long-lasting infrastructure that makes language resources and technology available and readily usable to scholars of all disciplines, in particular the humanities and social sciences, in an international context.

In this context, the repository's designated community are scholars of lexicography, general corpus linguists or typologists working with resources of a large number of languages. A special focus lies on scholars requiring resources for so-called "under-resourced" or "under-represented" languages and on resources generated via statistics-based methods and applications (statistical language models, topic models). Especially in the context of larger research infrastructure projects (like CLARIN or Text+) there is a strong development towards standardization of formats and interfaces which is actively promoted by the repository.

Links:

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Level of Curation

Select all relevant types of curation.

- Content distributed as deposited

- Basic curation – e.g., brief checking, addition of basic metadata or documentation

- Enhanced curation – e.g., conversion to new formats, enhancement of documentation
- Data-level curation – as above, but with additional editing of deposited data for accuracy

Compliance level:

Not Applicable - 0

Response:

- 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

Links:

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Level of Curation - explanation

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

Compliance level:

Not Applicable - 0

Response:

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 (which we would otherwise support, see below). 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, dissemination/discoverability and usability of the collection.

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

The repository encourages depositors to use standard formats and established vocabularies which allows including data without extensive curation effort. Especially in the case of descriptive metadata of resources, the repository supports creating and improving structured metadata records and additional documentation to be compliant with the repository's mission to be a valuable part of a modern and open distributed research infrastructure.

Information concerning data depositing and its criteria for support are also publicly available on the repository website (e.g. under https://repo.data.saw-leipzig.de/depositing/en).

Links:

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Insource/Outsource Partners

If applicable, please list them.

Compliance level:

Not Applicable - 0

Response:

Insource:
1) Saxon Academy of Sciences and Humanities in Leipzig (SAW)
The SAW as the hosting institution of the repository provides the basic infrastructure for maintaining and developing the repository. This includes administrative support, office space and similar necessities.

Outsource:
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). 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. A public document lists the services which are stipulated (http://www.clarin-d.de/mwiki/images/0/0b/GWDG_PID.pdf). The GWDG is also an active member of the Text+ consortium and provides its services (including this PID service) in its context.

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.net/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.
Part of this infrastructure will be taken over and continued by the Text+ project (of which the repository is a member of) in the future. This includes the helpdesk, monitoring service, and a new documentation platform.

CLARIN-D is represented by the German association „Geistes- und kulturwissenschaftliche Forschungsinfrastrukturen e.V.“ (http://www.textgrid-verein.de) both in national matters and in regard to CLARIN ERIC.

3) CLARIN-ERIC
CLARIN-D is a member of CLARIN's European Research Infrastructure Consortium (ERIC). 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 Observatory - VLO (https://vlo.clarin.eu): CLARIN's central metadata-based search engine which contains metadata of all German CLARIN-centres (among others).
– 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 (https://www.eudat.eu), access to advanced EUDAT services.

4) Text+

The repository is part of the Text+ consortium (started in autumn 2021). Text+ provides, to an increasing extent, services and infrastructural components (including a helpdesk, technical monitoring etc.) on which the repository will rely on. However, as Text+ is still in its starting phase the usage of these components will only increase over the coming months and years.

5) Leipzig University

Parts of the repository's infrastructure is hosted at Leipzig University. This includes two servers that are operated in server rooms of the university and which are administrated by repository personnel at SAW. The hosting situation is set out in an dedicated agreement between the repository and the responsible department of the university.

6) Other partners

- Verba Alpina
- SFB 1199
- OSIAN

More information about these institutions can be found under https://repo.data.saw-leipzig.de/partners/en . No binding agreements are yet in place, but affiliations with these partners are currently focused on consultation and a close, two-way communication concerning their (meta)data formats and depositing strategies, to plan and facilitate future collaborations.

Links:

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Significant Changes

Summary of Significant Changes Since Last Application if applicable.

Compliance level:

Not Applicable - 0

Response:

Since the last application, the repository was transferred from the Natural Language Processing Group (NLP group) at Leipzig University to the Saxon Academy of Sciences and Humanities in Leipzig (SAW). It is now hosted at an institution that is responsible for more than 20 long-term research projects in the humanities and that looks back on a particularly long tradition concerning the compilation of dictionaries and lexicographical resources. Thus, the transfer was a strategic move to ensure the sustainable operation of the repository on multiple levels. The SAW is legally a corporate body under public law, with the Free State of Saxony (Germany) as the sponsoring agency.

The transfer from Leipzig University to SAW was used to sharpen the thematic focus of the repository towards lexicographical resources and to update some of its core technical components including central parts of the archiving and backup infrastructure. This was done especially according to current requirements, to be a part of modern and developing research infrastructures. Furthermore, the definition and documentation of processes and workflows of the repository was improved (—> R12).

During the last CTS application, the centre was primarily funded via the CLARIN-D project, the German contribution to CLARIN. The funding of CLARIN-D has ended on August 31, 2020. However, CLARIN-D continues within the umbrella of the association „Geistes- und kulturwissenschaftliche Forschungsinfrastrukturen e.V." (http://www.textgrid-verein.de), and continues to be part of the European CLARIN family. The relevant CLARIN-D infrastructure will continue to be supported by and transferred to the Text+ project. This does not affect the European CLARIN infrastructure that is also

referenced in this document.

Links:

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Other Relevant Information

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

Compliance level:

Not Applicable - 0

Response:

The following requirements hold for CLARIN centres of type B:
– 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_checklist-v7_3_1.pdf).

These requirements are all fulfilled by this resource centre. Public evidence of its certification as a type B centre can be found here: https://centres.clarin.eu/centre/4

In part, similar criteria are currently developed in the context of the Text+ project (started in autumn 2021) which will be implemented by the repository in the future. This especially contains guidelines and policies to provide resources in a distributed federation of lexical resource centres. It furthermore means that the repository is embedded in a technical and organizational infrastructure that monitors state and quality of resources and services at the repository (e.g. using technical monitoring applications) and that supervises long-term development of the infrastructure as a whole and the contributions of each participating institution/repository (via administrative and scientific boards).

Links:

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Organizational Infrastructure

R1 Mission/Scope

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The mission of the CLARIN Resource Centre Leipzig at SAW 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. Public evidence of the mission statement can be found on the repository's main page [3].

The repository also serves as a repository of a CLARIN resource centre of type B ([4] – approved by the CLARIN coordinator's office). Public evidence of its certification as a type B centre can be found here [5].

This mission is supported by the infrastructure of the Saxon Academy of Sciences and Humanities in Leipzig (SAW) and by the integration of the repository into the national and international CLARIN infrastructures. As part of the CLARIN infrastructure, it shares the CLARIN mission to provide linguistic data, tools and services in an integrated, interoperable and scalable infrastructure for the Humanities and Social Sciences [1], and is committed to play an active role in the development of CLARIN's repository infrastructure. Similar goals are pursued in the context of the Text+ infrastructure that is still under development and that will play an important role in Germany’s federal scientific landscape regarding FAIR use of language resources, their long term storage, enabling their broad use in science and strengthening interoperability between data domains. The repository with its inventory of lexical and other linguistic resources participates in Text+ particularly in the data domain “Lexical resources”.

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 the German branch of CLARIN) 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. PDF: [2]

Links:

- 4 - About CLARIN
- 5 - CLARIN publication PDF
- 1 - Repository Landing Page
- 2 - Guide to CLARIN Centres
- 3 - CLARIN Centre Registry entry of the repository

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:

The repository only accepts data that is available under an open license (open data, publicly available). Resources in the repository are usually subject to the Creative Commons License CC BY-NC [5], if not stated differently on the description pages of respective resources. Currently, mostly data created by the hosting institution or affiliated projects is hosted at the repository. Depositors have to sign a depositor agreement [1]. 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

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 risks will be added to the repository, depositors also have to state that the deposited resource does not contain any data with a disclosure risk. The repository staff maintains a checklist of cases in which resources containing data with a disclosure risk have previously been rejected or modified by the depositor (and if so, how they were modified by the depositor) 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 risks.

In case a violation of conditions by users 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. Users are informed about the licenses in place on a designated page [2] and, if differing from this general information, on the description pages of particular resources.

The repository does not allow the integration of data without providing an appropriate license. These license conditions are available to users via CMDI metadata. In case of misuse, the only practical option is to deny the depositor 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 [3]. This document also includes information on how non compliance with the access permissions is handled.

Clear transparency of the repository's handling of IPR issues is also part of the requirements for CLARIN B centres. Public evidence of the repository's certification as such a type B centre can be found at [4].

Links:

- 7 - Repository Depositing Guidelines
- 8 - Repository Licensing Information
- 9 - Repository Access Permissions
- 3 - CLARIN Centre Registry entry of the repository
- 6 - Creative Commons BY-NC 4.0

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:

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 Saxon Academy of Sciences and Humanities in Leipzig (SAW) which is an institution governed by public law and which is co-financed by the Saxon State government out of the State budget approved by the Saxon State Parliament.

The SAW with its repository is a member of the Text+ consortium that is part of Germany's National Research Data Infrastructure (Nationale Forschungsdateninfrastruktur – NFDI) which is set up and funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF). All NFDI consortia are initially funded for 5 years; the current funding phase of the Text+ consortium runs from 2021 to 2026. However, the general plan of the NFDI is to be the long-term financed cornerstone of Germany's future research infrastructure.

Furthermore, the SAW is member of the “Geistes und kulturwissenschaftliche Forschungsinfrastrukturen e.V.” [1], which is a German registered association that pursues a sustainability strategy to promote the further development and networking of research infrastructures in the humanities and cultural studies in Germany and Europe.

The repository is also part of the European CLARIN research infrastructure, including its German branch CLARIN-D. CLARIN centres are set up as a distributed network, where each centre institution is a hub for the digital humanities and brings its own financial resources into CLARIN, which ensures continued availability. 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. Additionally, in case of a withdrawal of funding the repositories’ content would be transferred to another CLARIN centre as formulated in a Memorandum of Understanding by the centres of CLARIN-D [2]. This agreement is still in place, despite the changes to the repository host institution (the Memorandum was adapted accordingly) and this institutions memberships in Text+ and textgrid. The legal aspects of the process of relocating data to another institution is addressed by templates of license agreements provided in CLARIN. Depositors are informed that such relocations are possible under all agreed upon and while upholding all legal agreements.

Links:

- 10 - TextGrid e.V.
- 11 - CLARIN-D Memorandum of Understanding

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:

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 a disclosure risk. The repository staff maintains a checklist of cases in which resources containing data with disclosure risks 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 [1].

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 working groups, 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 immediately.

No data with a disclosure risk will be published in the repository, but compliance with disciplinary and ethical norms must still be ensured. In case of doubt, the depositor is contacted on the issue.

Implementation of the GÉANT Data Protection Code of Conduct (DP-CoC) is also part of the requirements for CLARIN B centres. Public evidence of the repository's certification as such a type B centre can be found at [2].

Links:

- 4 - Repository Access Permissions
- 3 - CLARIN Centre Registry entry of the repository

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 repository is hosted by the Saxon Academy of Sciences and Humanities (SAW) which can ensure long-term stability and sustainability as it is governed by public law and has a long history of successful long-term projects, with a focus on the humanities and a particularly long tradition concerning the compilation of dictionaries. With its various projects and available expertise in the respective scientific fields, it provides a stable institutional foundation for the repository. The Academy is directed by the Presidium or Board of Regents. It consists of the President, the Vice-President, the Secretaries of the three Academy Classes, and the Secretary-General. In accord with the research work of the Academy, the Plenum creates structural and project-oriented commissions. The specific project is lead by an dedicated principal investigator.
The general organisational structure is depicted in a publicly available organizational chart [4].

The SAW with its repository is a member of the Text+ consortium as a part of Germany's National Research Data Infrastructure (NFDI, [5]) which is setup and funded by the German Federal Ministry of Education and Research (BMBF). All NFDI consortia are initially funded for 5 years, but are intended to be part of a long-term national research infrastructure that extends beyond this initial funding period. The current funding phase of the Text+ consortium runs from 2021 to 2026. This allows to fund staff, IT resources and other relevant expenses (including travel expenses) for the given time period. Further details about the SAW, the repository's funding, and organizational contexts can be found in R3 Continuity of access.

The employed staff is highly qualified for the assigned tasks due to many years of experience in the operation and development of the repository and its interaction with its designated community. This includes the operation as a European CLARIN centre ("Service Providing Centre", "CLARIN B-centre") for more than 10 years with a thematic focus on the topics listed in R0.

The repository's staff has a broad knowledge base in the field of natural language processing and creation/provision of language resources, each having worked on different aspects of this field. All members of the repository have a background as computer scientists working with and researching on state-of-the-art technologies. With their long-time work on the project "Leipzig Corpora Collection" / "Wortschatz Leipzig" they are proficient in the computer linguistic domain, including topics like corpora and dictionary creation, language resource distribution, Web-based service interfaces, etc.

The work on related projects and cooperation with the Leipzig University allows the staff to further educate themselves as well as advance professionally, e.g. PhD theses. Workshops and joint knowledge exchanges and transfer in CLARIN and Text+ offers options for training and additional development ([1] -> "Learn & Exchange", [2], [3]). This includes active participation in relevant taskforces/working groups and direct contact with developers of applications relevant for the repository (like Lyrasis as developer of the repository's technical backbone Fedora).

At the moment, 5 staff members, including the project director, are working on the repository and its participation in the Text+ project, performing tasks like overall management, policy and software development, infrastructure design and planning, curation, ingestion of data and metadata, and contact with users and the designated community. Some roles are shared among staff to ensure redundancy in case of personal changes.

Links:

- 14 - CLARIN website
- 15 - Text+ Data and Competence Centres
- 16 - Text+ Cross-Cutting Topics
- 12 - Organigram
- 13 - NFDI website

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:

Advisors/Guidance:
The repository, through its membership in Text+ and CLARIN, is supported by several advisory boards and committees.The Text+ Scientific Board is the consortium's scientific lead and decides on the portfolio development. It is therefore a valuable guide for all questions regarding long-term development and questions regarding future interoperability with other projects and consortia [1]. Furthermore, the scientific expertise stemming from the former CLARIN-D technical advisory boards and discipline-specific working groups has been transferred into the newly established structures of Text+.

Text+ is structured along so-called "data domains" which are organized in thematic clusters.The repository is part of the data domain "Lexical resources" and the task area "Infrastructure/Operations". For each of these data domains exists a Scientific Coordination Committee (or Operations Coordination Committee) that evaluates and leads the scientific or operational development and provides feedback regarding topics like questions of technical protocols, infrastructural requirements on the level of archiving, interconnection, search, etc.. All of these committees are made up of established experts with many years of experience in their respective fields [2].
Besides these boards, there are participating researchers and developers in the various thematic clusters providing valuable feedback and guidance if needed.

Communicating with the Designated Community:
The repository's personnel is actively involved in scientific work in the field of lexicographical resources and technical aspects of large research infrastructures. Most staff members are therefore themselves part of our designated community, including participating in (or publishing at) domain-specific workshops and conferences. These opportunities are actively used to maintain and expand contact with our designated communities and to explore opportunities for improvement.
This effort also includes the participation in the CLARIN/Text+ help desk [5] which is established for many years now and which provides feedback/guidance for interested users for our offers but also for general questions about our areas of expertise. This feedback is also used to continuously adapt our offer to the needs of the community.

Other:
Text+ as a consortium of Germany's research infrastructure (NFDI) is in close contact to academic societies and other NFDI consortia. It cooperates with national and international associations in the field of language resources, services and the general topic of sustainable research infrastructures [3]. These contacts are also actively used to future-proof decision making and available if guidance and feedback is required.

The Saxon Academy of Sciences and Humanities (SAW) has long-term experiences with research projects in the humanities and provides a platform for academic-internal communication with researchers and research projects of scientific fields relevant for the repository. In addition, experts for scientific and operational topics are available for feedback and assistance. This includes a security officer for security- and safety related issues, a committee for linguistics [4], and a dedicated secretary (Referent) for the Digital Humanities. Regarding network and infrastructure security issues a dedicated security officer is also available at Leipzig university.

Furthermore, the repository's personnel is in active contact with Lyrasis, which develop the open source repository software Fedora, and its user community, allowing for technical guidance and assistance regarding the repository's underlying data storage software.

Links:

- 17 - Text+ Boards
- 18 - Text+ Coordination Committees
- 20 - Text+ Associations
- 21 - SAW Committee for Linguistics
- 19 - Text+ Help Desk

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 process of depositing and ingesting resources is documented as a BPMN 2.0 process model (see “Ingest” on [1]). As can be seen, there is close cooperation between the depositor and repository staff in case questions or problems arise in any stage of the process.

The depositor has to submit a defined set of documents along with the data:
- the signed Resource Deposition Request Form (RDRF)
- metadata corresponding to the submitted data
- an archive file containing the data and adhering to the BagIt format [2], which includes checksums for all contained files

If issues arise during the review of the RDRF document, the depositor will be asked to rectify the document within a defined period of time. After all RDRF issues have been resolved, the RDRF will be checked along with the data regarding several aspects, including the identity of the depositor, the origin of the data and licensing terms. In some cases the submission will be rejected 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. There are no separate checks concerning disclosure risks. Instead, depositors have to assure the nonexistence of such risks in the depositor agreement. If no more issues remain, the submitted metadata and the BagIt archive are validated and verified. In case the metadata validation is unsuccessful, revised metadata will be requested. If the BagIt format was not correctly implemented, which can be verified in an automated fashion with third-party software like the Library of Congress’ bagit-java [3], or if any submitted checksum contained in the BagIt archive does not match the checksum calculated for the contained data, we will ask the depositor to resubmit an updated BagIt archive.

After successfully passing all checks, the resource will be stored in a Fedora 6 repository system. During this process, 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. These mechanisms are closely linked to the storage procedures described in R9 such as the backup process (see “Backup” on [1]). At the end of the process, the depositor will receive a notification.

We adopted the BagIt format as the core structural convention for our digital objects as the format’s focus lies on strong integrity assurances in the context of storage and transfer of data and it is widely used for preserving digital assets by a significant number of organizations. In addition to the validation we perform upon ingestion, it can be re-verified again at any point, to ensure no accidental changes have been made to the data, including after dissemination, as downloads of the complete BagIt archives are provided in some cases.

Access to data and metadata is provided via different webservice interfaces, including an OAI-PMH endpoint [7]. The availability and standard conformity of these webservices is monitored via Icinga [8] probes. Some of these probes are run in a local installation at the centre while others are operated by the European CLARIN project [4] (requires academic sign-in) and (in the future) the German Text+ project. The frequency of checks depends on the type of service that is monitored.

Multiple versions of data are valid, especially as both CLARIN and Text+ propagate the idea of reproducible research. Thus, updates/new versions of existing data are 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 and other webservices
(2) a unique PID is assigned.

Thus, an update of already stored data files is not intended. Instead, a new resource, with a new PID, will be ingested. This ensures that the files of a resource will never be altered, and that alternative versions of the resource will be separately accessible and easily distinguishable.
As data updates are considered to be a new version of the existing data, data producers need to provide the same type and scale of information (metadata, documentation) that were provided for the previous version. A reference to a previous version is also required in the metadata.

Conversely, updates of existing metadata, which are ingested into the Fedora 6 repository alongside the BagIt archives, are allowed without considering the result to be a new resource. Instead, updates are managed via the Fedora 6 versioning system (utilising the Memento protocol [5]), allowing logging

and retrieval of all changes. Updated metadata are first parsed for syntactic correctness and manually evaluated for completeness and soundness and only then updated in the repository. They are also subject to the repository's fixity checking, ensuring their persistence. These fixity checks are carried out at regular intervals to ensure that unintentional changes are recognized and can be dealt with accordingly (like a "rollback" using backups).

The metadata provided for each resource that is to be added to the repository needs to contain basic information about 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 versions of a resource are present in the repository, references to previous/newer versions need 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 [6].

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 usability we encourage but not enforce data producers to provide change-logs if new versions of already existing data are ingested into the repository.

There is currently 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 or affiliated research projects. 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 possible legal issues.

Links:

- 22 - Repository Depositing Workflows
- 23 - BagIt File Packaging Format
- 24 - BagIt Library (BIL)
- 27 - CLARIN Monitoring
- 28 - HTTP Memento
- 29 - CLARIN FAQ CMDI Data Linking
- 25 - Open Archives Initiative Protocol for Metadata Harvesting
- 26 - Icinga Monitoring Software

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 guideline has been fully implemented in the repository - 4

Response:

The repository focuses on publicly available resources such as lexical data, language resources for so-called "under-resourced" languages and special reference corpora. If resources are submitted by data producers that do not fall within these criteria, the standard procedure is make the contact between data producer and a better fitting data repository in the context of the CLARIN or Text+ consortia.

The repository is closely related to the project Deutscher Wortschatz / Leipzig Corpora Collection (LCC) [1] and makes its resources available to various research communities. The repository's staff is also active in the development of the LCC, according to the Text+ working plan [2]. Therefore, just like the LCC, a main focus of the repository lies on general lexical resources, written text corpora, reference corpora, 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 [3]. More details can be found in R0.

For corpora and dictionaries of the Leipzig Corpora Collection, which are regularly 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)?
– Is the quality of the data high enough or can it be assessed at all?

The handling of requests to deposit data that does not fall within the CLARIN or Text+ mission will be decided on a case by case basis. Data that supports this 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, ideally following the BagIt convention, 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 are put into the repository are checked for compliance with internal and CLARIN/Text+ guidelines concerning scientific and scholarly quality. Only data that:
(1) are the result of research projects,
(2) come with exhaustive metadata,
(3) which's data structure are following established standards or are described by a sophisticated documentation (PDF/A),
(4) come with information on how the data was originally created
will be added to the repository. The data files themselves, the metadata and (if necessary) additional documentation are obligatory parts of each repository entry.

The previous criterion for third party review was dropped, as it was decided that well-documented, metadata-supported research data which underwent the aforementioned selection process was in and of itself sufficiently valuable to be added to the repository, even without third-party reviews (which in turn could, in some cases, result in very time-intensive organizational overhead).

These guidelines are available on the repository website [3]. They may be subject to minor changes in the future if the need arises.
The data stored in the repository is mostly well known and documented content as the result of long running research projects. For external resources we check the compliance 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 this repository has to be provided in the CMDI format (or at least in Dublin Core/DCMI as basis for CMDI records). There is exhaustive documentation [4] 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 [5] (public, PID present).
(3) Check if the data categories used in those components/profiles are present in the CLARIN Concept Registry [6].
(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 [7]. (5) Check if the provided CMDI files are considered of high quality according to the automatic analysis of CLARIN’s (metadata) “curation module” [8].

The granularity of CMDI metadata is up to the (meta)data producer. The repository itself is able to handle a high granularity of metadata, if necessary.
The creation of metadata records is supported via the Web-based editor COMEDI [9] 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 usable 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 [3]. Hints about mandatory or suggested metadata fields are also available [13].
In case the metadata provided is insufficient for long-term preservation, the data can not be accepted. In this case, we aim to support the depositor in providing the missing metadata.

It is recommended to use formats listed in the CLARIN standard recommendations
[10]. In addition, relevant standards and formats in the context of CLARIN are listed [11]. These lists and guidelines were generated in close collaboration with the respective communities of different fields of the humanities. Manual checks concerning the accordance with these guidelines are performed before data is added to the repository. Usage of standardized formats is encouraged but not enforced.
Similar recommendations and guidelines are currently worked on in the Text+ project. They will be encouraged/enforced as soon as they are approved and publicly available.

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 in 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 popularity.

If for any reason a resource needs to be removed from the repository, a placeholder “tombstone” will instead be associated with the existing persistent identifier, informing anyone accessing it about the resource’s new status. The PID and its connection to the resource will therefore be preserved, even if the underlying resource itself should not be available anymore. Rules on content removal, including accepted reasons to do so, can be found in the depositor agreement [12].

Links:

- 30 - Wortschatz Leipzig
- 31 - Text+ Participating Institutions
- 7 - Repository Depositing Guidelines
- 32 - CMDI Documentation
- 33 - CMDI Component Registry
- 34 - OpenSKOS
- 35 - CLARIN VLO
- 36 - CLARIN Curation Module
- 37 - COMEDI
- 39 - CLARIN Standard Recommendations
- 40 - CLARIN Standards and Formats
- 41 - Repository Depositor Agreement
- 38 - Repository Metadata Requirements

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:

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:

All relevant procedures and additional information for storage are documented in an internal MediaWiki instance, which allows for version-controlled management of all documentation, enabling the monitoring and handling of any changes to it. In addition, the preservation policy contains public information about data storage at [1].

Every relevant activity (ingest, storage, updating, data access, etc.) was tested by multiple people before moving to production and is documented in the MediaWiki instance. Storage locations and relevant hardware information are also documented there.

The underlying repository infrastructure software, Fedora 6, is compliant with the OAIS reference model, as are the repository's central workflows. Please refer to R12 for a more detailed breakdown.

The repository uses servers in the computing centres of the Saxon Academy of Sciences and Humanities and Leipzig University. All storage and backup solutions are set up redundantly. Maintenance of the systems is performed by the repository's system administrator. On the physical servers, virtual servers are hosted in the well-documented server virtualization management platform Proxmox. Server access is secured via firewall with fail2ban SSH hardening and IP range restrictions. Used hardware is monitored constantly (see below) and replaced when signs of hardware degradation are detected.

Data are stored on RAID systems and backups of the virtual machines are created on a regularly scheduled basis using the Proxmox backup mechanisms. This process is documented in a BPMN 2.0 diagram (see "Backup" on [2]). During the backup process, a VMA file and its checksum are created for the repository machines (see [5]). Once this backup has been stored, checksums are verified. If failing, any potential problems causing data alterations will be resolved and the backup will be created again. These backups are held on separate hardware, located in separate server rooms in a separate fire safety zone at Leipzig University and are regularly copied onto hardware of the Saxon Academy of Sciences and Humanities, thus strengthening storage location multiplicity. The VMA backups are performed weekly, while always preserving the five most recent versions.

In addition to the backups of the virtual machines, a second backup processes is put in place to increase redundancy and enable a fallback recovery option: Fedora 6 persists all of its data to an OCFL repository (Oxford Common File Layout [4]), located in a single directory. From the contents of this OCFL repository, the entirety of the repository (including a runtime database component) can be recreated. Backups of this OCFL repository are created by compressing a copy into an archival file, generating a checksum for verification and storing the backup files and their checksum logs on a separate physical machine. This allows for a complete recovery of the repository even in case the primary VM backup recovery fails. The OCFL backups are performed monthly, always preserving the three most recent versions.

In addition to the backups of recent VMA and OCFL versions, every six month an additional long-term backup is performed for both of them and stored alongside the renewed current backups, to allow the preservation (and, if needed, fallback) of older snapshots as well.

After an analysis of possible risks and failures, a recovery plan was drafted and added to the internal documentation, to ensure a precise sequence of steps can be implemented to re-instantiate data and/or systems upon different levels of failures, which were distinguished as:
1) Data loss/corruption
2) Software failures
3) Shutdown of hardware systems
4) Destruction of hardware systems

The integrity of the data elements is ensured via the BagIt convention [3], which stipulates the generation and storage of checksums for all its contained data files. Additionally, for backups of full virtual servers checksum tests are in place.

Deterioration of storage media is monitored via Icinga probes which perform regular checks 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 cases of failures/problems/etc. the administrators of the repository are notified and will take appropriate actions.

Each staff member is assigned workflows which are documented in a internal MediaWiki instance. Each staff member's roles and responsibilities are also documented there. In the context of the transferal of the repository to its new home institution (see Significant Changes in the Background Information section), role definitions and responsibilities have been streamlined and more definitely allocated. In addition, a new repository manager position was created to specifically oversee and continually improve all relevant procedures important to the repository, such as ingest, storage management and data access. All positions' responsibilities and their tasks' requirements were documented to be easily transferable in the future.

Links:

- 42 - Repository Preservation Policy
- 22 - Repository Depositing Workflows
- 23 - BagIt File Packaging Format
- 44 - OCFL
- 43 - VMA Format

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 guideline has been fully implemented in the repository - 4

Response:

All 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 (which will be subject to change based on future experience with depositors) is publicly available on the repository's website [1].

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.

It is up to the depositor to provide new versions of their respective data, including monitoring obsolescences of their chosen data formats.

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 (often based on the Handle system), moving resources from one CLARIN resource centre to another is possible without affecting the validity of references (e.g. PID of a resources used in a paper). This transfer to another CLARIN centre is formulated in a Memorandum of Understanding by the centres of CLARIN-D [3].

Our setup consists of virtual machines which may be moved to other partners. In case virtual machines are moved internally (inside the repository) 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 and established XML formats, etc.) 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 the very 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, including 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 (MariaDB, Apache Tomcat, Fedora 6 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 [2].

Links:

- 41 - Repository Depositor Agreement
- 42 - Repository Preservation Policy

- 11 - CLARIN-D Memorandum of Understanding

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:

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 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 trained 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 and Text+ projects, including those about metadata standardization and metadata quality.

The quality of resources is evaluated before any ingestion activities take place. This includes semi-automatic evaluation procedures based on resource-type-specific checklists and language-statistics based detection of anomalies and outliers in the text data (if applicable, see Thomas Eckart, Uwe Quasthoff and 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 concerning their conformity towards the applicable standards and them being well-formed by personnel having experience with the respective resource type. In case of problems, a re-submission of the adapted resources is possible, via the same procedures as the original submission. For resources provided by depositors, the depositors are encouraged to provide the repository with proof of the relevance of the resource and, if applicable, 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. 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 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 around 70 - 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 repositories based on a combined score of all evaluated features [4]. Further information about the curation module can be found at [5].

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 [1]
- easy-to-use Web interfaces to access sample data for manual inspection (e.g. [2])
- providing references to peer-reviewed publications about the data and their quality (if available)

Furthermore, the repository makes use 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). These records are accessible for end users in the faceted metadata search engine Virtual Language Observatory (VLO [3]) and (in the future) in the Text+ collection registry,
- 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 Help Desk (in the future: Text+ Help Desk), where help desk agents forward questions or remarks directly to the responsible centre/repository.

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

Links:

- 47 - Wortschatz Leipzig CLS Portal
- 48 - Wortschatz Leipzig Corpora Portal
- 35 - CLARIN VLO
- 45 - CLARIN Curation Dashboard
- 46 - CLARIN Metadata benchmarking and curation module

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:

Following this paragraph is a copy of the current version of the public documentation of the repository's central workflows, complying with the OAIS reference model, that can also be found at [6].
The web documentation will continually be updated if workflows are adapted and also includes more detailed BPMN 2.0 process models for the processes of ingesting resources and the VMA backup (see "Ingestion" and "Backup" on [6]), and links to additional documentation concerning (meta)data formats, data preservation and access permissions and privacy (among other topics).

The repository's general workflow is defined and implemented according to the OAIS reference model.
Following OAIS terminology, depositors take the role of the information producer, preparing a Submission Information Package (SIP) consisting (at the minimum) of:
– the signed Resource Deposition Request Form (RDRF)
– metadata corresponding to the submitted data
– an archive file containing the data and adhering to the BagIt format [7]

For more information on guidelines for resource depositing, including metadata requirements and data formats, please see [1] and [9].

The SIP will be verified in an appraisal phase, consisting of:
– a check of the RDRF
– verification of the metadata
– BagIt validation (including payload file checksums)

In case of discrepancies during any of these steps, a resubmission is necessary. If the appraisal was successful, the Archival Information Package (AIP), consisting of the BagIt archive file and the resource's metadata, will be prepared and ingested into the repository. This phase includes:
– checking for a persistent identifier (PID) present in the metadata

– if absent, registration of a new PID and an update of the metadata
– ingestion into the Fedora repository

Note that as no data with a disclosure risk will be published in the repository, no particular check for disclosure risks is needed here.

After storing the resource in the repository, the metadata is subject to regular fixity checks and, in case of updates, versioning. Updates of the payload files themselves will be treated as new resources and must be submitted in a new ingestion process. For more information on data preservation, please see [10].

Dissemination Information Packages (DIP) of the resource can be retrieved via various interfaces and applications, including:
– file downloads on the repository's Web page [2]. This is the default way to access resources of non-standard formats
– the OAI-PMH endpoint at [3]
– the CLARIN VLO [8]
– the CLARIN FCS (for resources with indexable full text, as the FCS is a full text search engine [4])
– various web services (depending on the resourcetype, e.g. web services for a dictionary-like retrieval can access dictionary resources, and so on)
For more information on access permissions and privacy, please see [5].

Any significant future changes to this workflow will need to adhere to the overall model structure of:
SIP —> Appraisal —> Ingest —> AIP —> Archival Storage —> Access —> DIP

They will need to ensure compliance with the presented (meta)data formats [1] and all relevant dissemination interfaces.
Only after these requirements have been confirmed and tested on a separate test instance of the repository will the workflow changes be implemented and documentation be updated.

Links:

- 7 - Repository Depositing Guidelines
- 49 - Repository Resources Overview
- 50 - Repository OAI-PMH Endpoint
- 51 - CLARIN FCS
- 9 - Repository Access Permissions
- 22 - Repository Depositing Workflows
- 23 - BagIt File Packaging Format
- 35 - CLARIN VLO
- 38 - Repository Metadata Requirements
- 42 - Repository Preservation Policy

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:

CLARIN Center Leipzig

All CLARIN centres [7] provide their metadata in the CMDI format. The Component MetaData Infrastructure (CMDI, ISO 24622, [2]) 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 [3].

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 [8]). 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 Leipzig repository, is to make metadata for all resources available through the established Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [1]. This standard enables harvesting of the metadata from the repository via http(s).The repository's OAI-PMH endpoint can be accessed at [6].

Search facilities are currently not provided by the repository itself. Instead, services of the CLARIN project 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 at [9].

Metadata:
The CLARIN Virtual Language Observatory (VLO) [4] harvests metadata in CMDI format from all CLARIN centres via OAI-PMH, including our metadata. Metadata from CLARIN centres (and other relevant archives and repositories) are browsable and searchable via this Web search portal. 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. The VLO then provides information pages about the resources, including licenses, and links to the repository for direct downloads of the files.

The quality of metadata is continuously assessed in CLARIN regarding, among other things, the presence of metadata entries that are of particular importance for user searches [10].

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 [11] interface. The repository also offers this kind of access for some of its resources, specifically for the corpora of the Leipzig Corpora Collection.

In addition, the repository offers its own portal for lexical searches in the datasets of the Leipzig Corpora Collection. It is available at [5].

PIDs:
The repository uses the common CLARIN PID service [12] based on the Handle System [58] 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. PIDs are easily accessible both on every resource's page in the repository Web frontend and the VLO.
As an example, for each resource of the Leipzig Corpora Collection, which is made available in the CLARIN context, PIDs using the Handle System are available, e.g. for the text corpus "nld_news_2012_300K": [14].

Text+:
During the next years, the German research infrastructure project Text+ will build an infrastructure with a partly similar focus as the CLARIN infrastructure focusing on the German scientific context. The Text+ infrastructure will provide applications for an efficient discovery of data and metadata as well, including central registries for data and services. The Leipzig repository is participating in the development of these applications and services and will integrate its data inventory in it.

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 repository and the VLO (as well as in the metadata records). E.g. for the corpus "nld_news_2012_300K":
- Repository-Frontend: [15]
- VLO: [16]
- VLO: [17]
Additional information on recommended data citations is also available via our corpus search portal (e.g. [18] → "Details").

Links:

- 32 - CMDI Documentation
- 33 - CMDI Component Registry
- 35 - CLARIN VLO
- 48 - Wortschatz Leipzig Corpora Portal
- 50 - Repository OAI-PMH Endpoint
- 52 - CLARIN Centres Overview
- 53 - CLARIN Concept Registry
- 54 - CLARIN Value Proposition
- 55 - CLARIN Collection Report SAW Leipzig Repository
- 56 - CLARIN FCS Documentation
- 57 - CLARIN PID Guide

- 58 - Handle System
- 59 - Example Handle
- 25 - Open Archives Initiative Protocol for Metadata Harvesting
- 60 - Repository Frontend Citation
- 61 - VLO Citation Example
- 62 - VLO Citation Example, Additional Information
- 63 - Corpora Portal Citation Example

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 guideline has been fully implemented in the repository - 4

Response:

Metadata for all CLARIN repositories has to be provided in the CMDI format. There is exhaustive documentation [3] 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.

Metadata records have to be publicly provided via an OAI-PMH endpoint [2] under an open-source license or without any legal restrictions (public domain). The Leipzig repository provides there metadata records as part of the public domain (documented here [1]) Since providing Dublin Core metadata is part of the OAI-PMH standard, those are also available for all the repository's resources.

Resources must be accompanied with valid CMDI metadata in order to be considered for deposit. As an alternative, Dublin Core metadata can be provided which will be used to create CMDI metadata instances. 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 [4] (public, PID present).
3. Check if the data categories used in those components/profiles are present in the CLARIN Concept Registry [5].
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 [6] by using CLARIN's "curation module" Web application (available at [7]).

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 at [9].

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_1527668176047)
OLAC-DcmiTerms(clarin.eu:cr1:p_1288172614026)
singlePaperPackage (clarin.eu:cr1:p_137588037297)

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 Leipzig repository (like the aforementioned 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 [10]. In addition relevant standards and formats in the context of CLARIN are listed [11]. These lists were generated in close collaboration with the respective communities of different fields of the humanities. Usage of standardized formats is encouraged but not enforced. For standardized formats, the metadata record format's validity is checked manually by repository staff before ingest.

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's maintainers keep track of all formats already used by the depositors and commit themselves to work on updates of the CLARIN standard recommendations if new formats gain in popularity. This is currently done in conjunction with CLARIN's standards committee [8]. 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.

During the next years, the German research infrastructure project Text+ will build an infrastructure with a partly similar focus as the CLARIN infrastructure focusing on the German scientific context. The Text+ infrastructure will provide applications for an efficient reuse of data and metadata as well, including central registries for data and services. The Leipzig repository is participating in the development of these applications and services and will integrate its data inventory into it. The repository's personnel is already actively working on standard recommendations in the Text+ context.

Links:

- 8 - Repository Licensing Information
- 50 - Repository OAI-PMH Endpoint
- 32 - CMDI Documentation
- 33 - CMDI Component Registry
- 34 - OpenSKOS
- 35 - CLARIN VLO
- 36 - CLARIN Curation Module
- 64 - CLARIN Standards Committee
- 38 - Repository Metadata Requirements
- 39 - CLARIN Standard Recommendations
- 40 - CLARIN Standards and Formats

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:

Technology

R15 Technical infrastructure

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The repository is using well-supported LTS operating systems and additional core infrastructural software which is appropriate to the services it provides to its designated community and which is available under open source licenses. The repository has a strong focus on open-source applications and public standards. Proprietary software or extensions are not used to avoid vendor lock-in or insufficient update procedures.

The servers are located in modern server rooms, which provide air conditioning and (limited) uninterrupted power supplies. Access to the server rooms is limited to authorized staff.

Debian Linux [1] is used as the operating system and Proxmox VE [2], an open-source platform for enterprise virtualization, is utilized for virtualization. The repository itself is based on the Fedora 6 repository solution [5][6] in combination with other popular open-source applications, like MariaDB [7] and Apache Tomcat [8].

The services of the repository are running on three main physical servers:
A) a production system with a Proxmox VE host: all service/websites/webapps run on dedicated virtual systems
B) a backup production system with a Proxmox VE host to allow for quicker recovery of the repository in case of a failure of the main host
C) a data backup system: Physical server with over 100TB for offside backups (including VM-snapshots, database-dumps, OCFL directory)
All physical servers are equipped with RAID systems.
The production system includes two dedicated virtual machines for different tasks (base repository, dissemination services). For each VM weekly snapshots are created (last five snapshots will be stored) and stored on the data backup system. All VMs are running on the Debian operation system with several protection mechanisms like restricted IP ranges for SSH/database access, fail2ban and firewall. 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 continuously performs checks for network connection, system load (LOAD), memory usage (MEM), disk space usage (HDD) and update status. This Icinga instance is running independently from and in addition to the central CLARIN monitoring (also based on Icinga). The CLARIN monitoring (in the future: Text+ monitoring) is mainly used for availability checks and for checking the standard compliance of provided interfaces (like OAI-PMH) and is hosted externally, at another CLARIN centre.

To guarantee that all software components are up-to-date the repository's local monitoring checks if system updates are available, which are then installed by technical personnel. The workload and remaining capacity of all VMs is monitored, to react in case of an overload situation. This could e.g. be caused by massive automatic querying of services, which can then be stopped by banning of IP addresses (or IP address ranges), or warnings in case of low disk space.

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. A similar procedure is carried out to evaluate the appropriateness and update status of the complete software stack in use and all of its dependencies. This is done in addition to system update checks based on Icinga.

The repository uses additional internal applications which are hosted by its institution (SAW). This includes a code repository based on Gitlab [9] and a MediaWiki instance for documentation [10]. All of these applications are popular, well supported open source applications.

Internal system documentation is available to all staff members in an internal MediaWiki instance. This documentation is constantly 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
- setup guides (restoring backups, recreating software and services from existing configurations)
- semantic annotations for structured metadata querying and filtering (e.g. overview of and relationships between services and servers, installed resources)

For creating and providing metadata we rely on the standards of the "Component Metadata Initiative" (CMDI, ISO-CD 24622-1). CMDI records (and Dublin Core records) are available via the standard interface OAI-PMH [3]. The CMDI file of a resource contains links to documents stored in the repository, interfaces – usually web services – or web applications that facilitate usage of the resource.
The repository complies with the OAIS reference model's tasks and functions. It is based on a modern Fedora repository instance, which is compliant with the OAIS reference model and a popular platform with an active community and developers [5].

Within CLARIN and Text+, standardization and use of standards is discussed and reviewed on a regular basis. This is promoted in dedicated committees like the CLARIN Standards Committee [4]. As part of CLARIN and Text+ we are committed to play an active role in the development of a distributed repository infrastructure. General plans for maintaining and further developing the infrastructure have been formulated as part of the project proposal or work plans.

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 to monitor advances in infrastructure development take place regularly.

The technical infrastructure is designed to be quickly recoverable even in the event of major hardware outages. VMA backups can be moved to new hardware if necessary, or alternatively, a new virtual machine can be created via our virtualization platform on which the repository can then be freshly initialized using the OCFL backups.

Links:

- 65 - Debian Linux
- 66 - Proxmox
- 25 - Open Archives Initiative Protocol for Metadata Harvesting
- 64 - CLARIN Standards Committee
- 67 - Fedora
- 68 - Fedora Github
- 69 - MariaDB
- 70 - Apache Tomcat
- 71 - Gitlab
- 72 - MediaWiki

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 repository is in the implementation phase - 3

Response:

In the course of the transferal of the repository to its new home institution (see Significant Changes in the Background Information section), security measures and in-place safety systems were re-evaluated and extended, especially concerning allocation of access rights and general network access possibilities (see below). For details on the repository's continuity plan, please refer to R3.

The repository is only being used to store open-access datasets (e.g. publicly available and accessible language corpora or dictionaries). There is no risk of disclosure for any data that might be leaked or accessed in a non-authorized way.

The physical servers for hosting the repository are in non-publicly accessible spaces, with access limited to authorized personnel only (Computing & IT Service staff). The server locations are secured with key and passcodes and have an active alarm system if no one is on-site. Locations have dedicated security officers for facility security as well as patrolling officers. The server rooms are equipped with climate control and fire extinguishing systems and are located outside of flood-risk zones.

Systems for managing source code and documentation use standard log-in mechanisms (e.g. LDAP, Shibboleth). Server access is secured with IP range restrictions, fail2ban SSH hardening, and credentials require industry standard minimum requirements for password strengths. Standard and established authentication methods are being used. Software and services are restricted to the minimum access levels required, e.g. public websites with read-only access to databases. Staff authorization for software and hardware is role-based and by responsibility only. Furthermore, Leipzig University is currently working on upgrading its firewall systems with a DMZ ("demilitarized zone"), which will greatly increase access security. This upgrade takes place in close communication with the hosting institutions, including this repository. We plan to apply for level 4 for this requirement in the next certification period, when the DMZ will be in place.

The SAW employs a monitoring system (Icinga) to monitor IT hardware and services. These systems inform the system management staff in case of critical issues or outdated software. Staff is assigned to monitor software status and execute updates if necessary. This is based on LTS releases of operating systems and other standard software applications and their available updates and bugfixes. In addition, the repository is integrated into the joint CLARIN technical monitoring (also based on Icinga) which is used to evaluate availability and standard conformity of public services. A similar monitoring system is currently built in the context of the Text+ project with a comparable focus. In addition, every six months the computing center of the Leipzig

University prepares a security assessment report based on a weak-point analysis of the complete network structure. This is concomitant with a demand to quickly resolve all such identified security flaws, ensuring a monitored, secure network infrastructure.

Links:

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:

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:

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.

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: