Assessment Information

CoreTrustSeal Requirements 2020–2023

Repository:	Tübingen CLARIN-D Repository
Website:	http://www.sfs.uni-tuebingen.de/ascl/clarin-center/repository.html
Certification period:	01 June 2023 - 31 May 2026
Requirements version:	CoreTrustSeal Requirements 2020-2022
This repository is owned by:	**University of Tübingen**

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background Information

Repository Type

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

Compliance level:

Not Applicable - 0

Response:

- Domain or subject-based repository

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Description of Repository

Provide a short overview of the repository.

Compliance level:

Not Applicable - 0

Response:

The Tübingen CLARIN-D repository (https://uni-tuebingen.de/en/134314) is a member of the European CLARIN research infrastructure and the research infrastructure consortium Text+, which in turn 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.

Among the resources currently available in this repository, researchers can find widely used treebanks of German (e.g. TüBa-D/Z), the German wordnet (GermaNet), the first manually annotated digital treebank (Index Thomisticus), as well as descriptions of tools used by the WebLicht execution engine for natural language processing.

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+ clusters "Lexical resources" and "Collections".

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Clear and concise

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 an infrastructure that makes language resources and technology available and readily usable to scholars of all disciplines, in particular the humanities and social sciences.

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

In 2012, nine CLARIN member countries created CLARIN-ERIC (European Research Infrastructure Consortium), which is an international legal entity that governs and coordinates CLARIN activities. CLARIN-ERIC members are governments or intergovernmental organisations which pay an annual fee to support the development and maintenance of the CLARIN research infrastructure. Germany is one of the founding members of CLARIN-ERIC and has contributed to CLARIN-ERIC via CLARIN-D (https://www.clarin-d.net/en/). CLARIN-D is an acronym for “Common Language Resources and Technology Infrastructure Deutschland”, the German national node of CLARIN.

The CLARIN Resource Center Tübingen is one of currently eight German CLARIN Resource and Service Centers which form a web and centers-based research infrastructure for the humanities and social sciences. The aim of CLARIN and its service centres is to provide language data, tools and services in an integrated, interoperable and scalable infrastructure for researchers in the humanities and social sciences and related disciplines. The research infrastructure has been rolled out in close collaboration with expert scholars in the humanities and social sciences, to ensure that it meets the needs of users in a systematic and easily accessible way. The CLARIN Resource Centre Tübingen is part of the CLARIN-D consortium funded by the German Federal Ministry for Education and Research. This funding ended on August 31, 2020.

CLARIN-D built on the achievements of the preparatory phase of the European CLARIN initiative as well as CLARIN-D's Germany-specific predecessor project D-SPIN. These previous projects have developed research standards to be met by the CLARIN service centres, technical standards and solutions for key functions, a set of requirements which participants have to provide, as well as plans for the sustainable provision of tools and data and their

long-term archiving.

Within CLARIN, this resource centre is a certified centre of type B, see Relevant Information below.

Since autumn 2021, the repository is member of the Text+ consortium, which shares CLARIN's mission and whose other members are either CLARIN members or closely affiliated with the CLARIN community. Compared to CLARIN, Text+ significantly increases the community to additional fields of the humanities; it also serves as a bridge to NFDI communities interested in textual and language-related data.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Clear statement of the community and the connections to others serving that community

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Level of Curation

Select all relevant types of curation.

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

Compliance level:

Not Applicable - 0

Response:

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

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

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:

This repository distributes data as deposited. However, prior to ingestion into the repository, data is checked for suitability for inclusion, and extensive metadata are created.
Among the resources currently available in this repository, researchers can find widely used treebanks of German (e.g. TüBa-D/Z), the German wordnet (GermaNet), the first manually annotated digital treebank (Index Thomisticus), as well as descriptions of tools used by the WebLicht execution engine for natural language processing.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Concur that this meets Basic Curation (B)

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Insource/Outsource Partners

If applicable, please list them.

Compliance level:

Not Applicable - 0

Response:

1) Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
The repository makes use of a common CLARIN PID service (https://www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf) based on the Handle System (http://www.handle.net/) and in cooperation with the European Persistent Identifier Consortium (EPIC). The usage of PIDs is mandatory for resources in CLARIN thus all resources added to the repository may be referenced using PIDs.

CLARIN-D has a contractual relationship with GWDG concerning the provision of PID-services via EPIC API v2. The following document lists the services which are stipulated: http://www.clarin-d.de/mwiki/images/0/0b/GWDG_PID.pdf

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

CLARIN-D offers several services to its member institutions, among them the following:

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

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

3) CLARIN-ERIC
CLARIN-D (https://www.clarin-d.net/de/) 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 Obervatory - VLO (https://vlo.clarin.eu): CLARIN's central metadata-based search engine, which contains the metadata of all German CLARIN-centres.
- Metadata harvester: The VLO is kept up to date using the metadata harvester run by the CLARIN- ERIC.
- Federated Content Search - FCS (https://www.clarin.eu/contentsearch): Optionally, centres can provide the actual data of their resources for this central content search.
- CMDI Component Registry (https://catalog.clarin.eu/ds/ComponentRegistry): CLARIN's registry for components and profiles according to ISO-24622-1.

In addition, CLARIN-ERIC offers several further services such as central registries (e.g., terminology), user statistics management and, as an official EUDAT community, 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.

Text+ is a consortium of the initiative to establish a national research data infrastructure (Nationale Forschungsdateninfrastruktur, NFDI). Text+ officially started in autumn 2021 after several years of preparation and will initially be funded for five years by the German Research Foundation. 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 aim of the national research data infrastructure (NFDI) is to systematically manage scientific and research data, provide long-term data storage, backup and accessibility, and network the data both nationally and internationally. The NFDI will bring multiple stakeholders together in a coordinated network of consortia tasked with providing science-driven data services to research communities.

5) University of Tübingen
Parts of the repository's infrastructure is hosted at the University of Tübingen. This includes a server that is operated in server rooms of the university's computing centre and which, in part, is administrated by repository personnel in the computing centre.

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Clear statement of partnerships No concerns

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:

-

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Other Relevant Information

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

Compliance level:

Not Applicable - 0

Response:

Within CLARIN, this resource centre is a certified centre of type B. CLARIN distinguishes a number of different centre types that have different impact for the language resources and tools infrastructure. Type B centres offer services that include the access to the resources stored by them and tools deployed at the centre via specified and CLARIN compliant interfaces in a stable and persistent way.

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

- Centres need to offer useful services to the CLARIN community.
- Each centre needs to refer to CLARIN in a visible way on its website.
- Each centre needs to make explicit statements about its funding support state and its perspectives in this respect.
- Each centre needs to make explicit statements about CLARIN compliant resources and services available at the centre.
- Each centre needs to make clear statements about their policy of offering data and services and their treatment of IPR issues.
- The centre has to implement the GÉANT Data Protection Code of Conduct (DP-CoC) for each of its federated Service Providers.
- Centres need to have a proper and clearly specified repository system and participate in a quality assessment procedure as proposed by the CoreTrustSeal.
- Centres need to adhere to the security guidelines and provide a reference to those guidelines.
- 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).

In part, similar criteria are currently developed in the context of the Text+ consortium (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 the state and quality of resources and services in the repository (e.g. using technical monitoring applications) and that supervises the long-term development of the infrastructure as a whole and the contributions of each participating institution/repository (via administrative and scientific boards).

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

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 Tübingen CLARIN-D Repository is to ensure the availability and long-term preservation of resources in the field of Humanities and Social Sciences, to preserve the knowledge gained in research, to aid the transfer of knowledge into new contexts, and to integrate new methods and resources into university curricula.

This mission is supported by the infrastructure of the University of Tübingen 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 (https://www.clarin-d.net/en/about), and is committed to play an active role in the development of CLARIN's repository infrastructure.

The CLARIN center in Tübingen supports data from the Humanities and Social Sciences with a clear emphasis on language related material, both for disciplines working with language analysis as the objective of research and as a research method. This covers data especially from Linguistics, Psycholinguistics, Corpus Linguistics, Syntax, Semantics, Lexicography, etc. but also includes other areas such as literary studies, political sciences, history etc.

The mission of the Tübingen CLARIN-D Repository is in line with the mission and goals of the CLARIN research infrastructure and Text+, namely, to provide easy and sustainable access for scholars in the humanities and social sciences (HSS) to digital language data (in written, spoken, video or multimodal form) and advanced tools to discover, explore, exploit, annotate, analyse or combine them, independent of where they are located.

The Tübingen CLARIN Repository is a certified Type B CLARIN Center (https://www.clarin.eu/content/eberhard-karls-universität-tübingen).

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

I concur in the assessment and have no suggestions.

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

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:

Before data can be deposited into the data repository, the depositor must provide an appropriate End User License Agreement (EULA) and sign a Depositor's Agreement.

The End User License Agreement (EULA) is an agreement between the depositor and the user, is provided by the depositor, and must be accepted by the user before obtaining the data. The repository encourages depositors to place data under open licenses such as Creative Commons whenever possible. Information about a dataset's licensing is stored in the CMDI metadata. In case of misuse of data, legal action may be taken by the depositor or property rights owner.

The Depositor's Agreement is an agreement between the depositor and the repository, where the repository is represented by the University of Tübingen. The Depositor's Agreement includes granting distribution rights to the repository, specifying access rights to the data (public, academic, individual), and assures that IPR and privacy rights are respected in the deposited data. The data provider retains all intellectual property rights to their data.

Access rights are enforced technically and the repository can restrict the downloading of a dataset to certain individuals or to the academic community. For some resources (e.g. those with individual access), the user may need to sign a license agreement with the depositor before the repository can give access to the resource. This is the case for datasets which are based on copyrighted material, such as newspaper text, and a license is required to protect the rights of the copyright holder. In this case, credentials are provided to the individual, which allows them to download the dataset. Access to resources that are limited to academic use is protected via Shibboleth and is only available to those that are able to login through IDPs operated at institutions taking part in the DFN-AAI or similar AAI federations that are part of CLARIN. This currently includes over 2,000 academic institutions in Europe.

If a problem is discovered in a dataset (such as personal data disclosure), the depositor will be contacted immediately, and steps will be taken to ensure that the data is not distributed until the issue is resolved. These steps include blocking download access to the dataset and preventing the metadata from being harvested. Additionally, records are kept by the repository for those resources which require the user to sign a license before gaining access credentials. These records of licensed users enables the repository to contact users in case of a legal issue with a specific version of a dataset, for example in case of legal action against a depositor from a third party. Users can also be informed if a new version of the dataset, in which legal conflicts have been removed, becomes available. The records of users are kept private according to German privacy legislation.

Please see the repository agreements and guidelines for more information (https://uni-tuebingen.de/en/134320).

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: I concur in the assessment and have no suggestions

Reviewer 1:

Compliance level:

Not Applicable - 0

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:

As stated in the Depositor's Agreement (https://uni-tuebingen.de/en/134320), the repository ensures that the deposited data will remain archived in a legible, accessible, and sustainable manner to the best of its ability and resources.

The repository is 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 (https://www.clarin-d.net/en/about/centres/mou-taking-other-centre-s-data). The legal aspects of the process of relocating data to another institution is addressed by templates of license agreements provided in CLARIN.

The repository is 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 repository is a member of the "Geistes und kulturwissenschaftliche Forschungsinfrastrukturen e.V." (http://www.textgrid-verein.de/), 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.

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

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:

The process of depositing data into the repository at the CLARIN-D center in Tübingen includes steps to ensure that the legal and ethical requirements of the archive are met. This is achieved organizationally through a contractual agreement (Depositor's Agreement) between the depositor and the archive, and also through close collaboration with the depositor prior to ingesting the data into the repository.

The Depositor's Agreement explicitly states that the depositor has fulfilled ethical and legal obligations, has resolved any disclosure risks related to privacy issues, and has respected Intellectual Property Rights (IPR) with regard to the data. In particular, data must be anonymized when applicable. Data depositors are responsible for compliance with all relevant national or international legal regulations. The depositor can choose to make the data publicly available, restrict access to academics via AAI (Authentication and Authorization Infrastructure), or to restrict access to individual users (which may be necessary if the dataset is based on copyrighted material such as newspaper text).

Only data that is in compliance with the Depositor's Agreement and with the University of Tübingen's guidelines for safeguarding good scientific practice (https://uni-tuebingen.de/en/119123) will be considered for depositing. In addition, we ask the depositor whether the data to be deposited contains any parts that have been contributed by third parties, and thus constitute a potential disclosure risk. If so, written documentation that the third party has

consented to redistribution of the data must be provided by the depositor.

Neither the CLARIN-D resource center, nor the repository run by it, are legal entities on their own. This also holds for the General and Computational Linguistics Department ("Seminar für Sprachwissenschaft", SfS) where the CLARIN-D center and its repository are located. All are part of the University of Tübingen, which is a legal entity - specifically, like all public German universities, a Körperschaft des öffentlichen Rechts, an institution governed under public law. Hence, the university as an institution is the contractual party in the depositor's agreement with appropriate authorities signing the contract.

The Depositor Agreement is governed by German law, hence the authoritative version is in German. An informative version in the English translation has also been made. Both are available here: https://uni-tuebingen.de/en/134320. For legal reasons, these agreements are templates only, to be adjusted on a case-by-case basis. A Depositor's agreement must be signed prior to depositing data.

The repository encourages users, depositors, and researchers to report violations. The repository has created an email address for this purpose; also, a user-friendly technical means for reporting violations have been recently added. In the case where the repository discovers that the Depositor's agreement has been violoated, distribution of the data and metadata will be halted, with the possibility of ingesting a new version in which the problematic parts have been removed. In the case of violation of the EULA by a user, further access by that user may be blocked. In all cases, the original data provider will be contacted.

As member of the Text+ consortium, the repository team is part of the transdisciplinary working group on legal and ethical issues within the NFDI context to monitor and implement all relevant developments on legislation (GDPR, ...). This working group also addresses ethical and related aspects of research data management.

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Extremely clear documentation and demonstration of meeting the guideline

Reviewer 1:

Compliance level:

Not Applicable - 0

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:

This repository is hosted by the CLARIN center Tübingen within the General and Computational Linguistics Department of the University of Tübingen. The center currently has seven staff members, four of whom are responsible for CLARIN technical activities, including operation of the repository. The repository staff have both the technical expertise and knowledge of language data required to ensure the safety of the repository and the quality of its data.

Neither the CLARIN-D resource center, nor the repository run by it, are legal entities on their own. This also holds for the General and Computational Linguistics Department ("Seminar für Sprachwissenschaft", SfS) where the CLARIN-D center and its repository are located. All are part of the University of Tübingen, which is a legal entity - specifically, like all public German universities, a Körperschaft des öffentlichen Rechts, an institution governed under

public law with basic funding provided by the Land of Baden-Württemberg.

CLARIN centers are hosted by scientific institutions - their repository staff members have access to training on data management, metadata, long-term preservation and professional development (offered by CLARIN-ERIC). This includes regular developer meetings, mobility grants for sharing of expertise, conferences, meetings with their respective scientific communities (called discipline-specific working groups) as well as a centralized knowledge base (user guide, wiki, bugtracker and mailing lists). CLARIN has a wide field of expertise in its collaborative network of centers, which come from within their respective fields of digital humanities. As part of CLARIN, staff members also have access to information on a wide range of topics that CLARIN offers, including teaching material, help on data management plans and other, discipline-specific support.

The repository is a member of the Text+ consortium as a part of Germany's National Research Data Infrastructure (NFDI, https://www.nfdi.de/?lang=en) 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.

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 many 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 or (computational) linguists working with and researching on state-of-the-art technologies. With their long-time work on GermaNet, treebanks, natural language processing tools etc. they are proficient in the computer linguistic domain, including topics like the creation of lexical resources and corpora, the distribution of language resources, the design and implementation of Web-based service interfaces etc.

Workshops and joint knowledge exchanges and transfer in CLARIN and Text+ offer options for training and additional development (https://www.clarin.eu -> “Learn & Exchange”, https://www.text-plus.org/en/research-data/data-and-competence-centres/, https://www.text-plus.org/en/networking/cross-cutting-topics-2/). This includes active participation in relevant taskforces/working groups and direct contact with developers of applications relevant for the repository.

The Tübingen CLARIN center is also supported by the University of Tübingen as part of the commitment of the university to support Text+. As a result it secures two full staff positions ("Eigenmittel", 2 FTE).

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R6 Expert guidance

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The repository, through its membership in Text+ and CLARIN-D, 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 (https://www.text-plus.org/en/about-us/boards-2/).

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 "Collections". For each of those 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 (https://www.text-plus.org/en/about-us/coordination-committees-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 lexica (e.g., GermaNet), corpora (e.g, Tübingen Treebank of Written German - TüBa-D/Z), tools (e.g., WebLicht pipelines for linguistic annotation, the Language Resource Switchboard) 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, 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.

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 (https://www.text-plus.org/en/networking/research-network/). These contacts are also actively used to future-proof decision making, and they are available if guidance and feedback is required.

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

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 integrity and quality of digital objects are ensured by two processes: A manual process at the time of ingestion, and an automatic process for continual monitoring of integrity. The identity of data providers is known and a Depositor's Agreement must be signed before data is archived.

Prior to ingest, an initial quality check and assessment of the provided data and metadata is performed by the archive manager. This involves checking the consistency and validity of the metadata and a check of the data provided. The metadata is validated for syntactic correctness and manually evaluated for completeness and soundness. Software tools (both standard and custom-built) support the archive manager in this process, for example in checking the validity of an XML document against its schema, or checking metadata for consistency. Any issues that arise can be resolved between the depositor and the data manager prior to ingest.

Bagman (https://weblicht.sfs.uni-tuebingen.de/bagman/), a web-based software, has been recently developed and deployed to support researchers in submitting their research data to the archive. Metadata is collected via forms and the entire package (research data and associated metadata) is then submitted using the BagIt format, see https://datatracker.ietf.org/doc/html/rfc8493. We have started using Bagman for researchers located at the University of Tübingen, and initial feedback is very encouraging. Once the bag has been received by the archive, a Python-based script is invoked allowing the archive manager to automatically check the completeness (all files have been transferred) and correctness (the checksum of each file transferred matches the checkum recorded in the bag) of the data transfer. The method complements the workflow described in the previous paragraph in cases users use Bagman for archival. Bagman, hence, supports both researchers and archive managers.

The integrity of the data is ensured by the version control in the Fedora-Commons backend. Metadata is a data stream within the digital object, and as such is version-controlled like object data. The system performs integrity checks of the individual data streams based on MD5 and SHA256 checksums. Problems and changes of files are reported to the archive manager for immediate action and restoring from backup if necessary.

The repository stores data but does not process or alter it in any way. Alterations of primary data is not allowed, but new versions of the data may be made available. New versions are assigned a version number and are stored in a separate data stream, which has an associated checksum which is automatically computed by the repository. In the rare case where a datastream needs to be updated, the previous version is automatically maintained by a version control system built into the repository back end. Metadata may be updated if need arises by the data depositor or the archive manager (e.g. to update contact information, addresses, or to add descriptions in other languages). The repository includes an RDF store of system metadata and also an audit trail which can be used to inspect past activities.

Metadata according to ISO 24622-1 (CMDI; http://www.clarin.eu/cmdi/) is uploaded or created during the archiving process. This step is required in the uploading process, since data without metadata is technically not accepted in the system. The front-end of the archiving system includes software to assist the depositor in creating valid CMDI metadata using components and profiles stored in the Component Registry (http://catalog.clarin.eu/ds/ComponentRegistry/).

To support interoperability with other metadata standards, the repository team offers automated transformation scripts into MARC21, DublinCore, HTML, RDF, and now also JSON-LD.

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Concur Clear documentation and no comments

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R8 Appraisal

Tübingen CLARIN-D Repository

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:

This repository supports data from the humanities and social sciences with a clear emphasis on language related material, both for disciplines working with language analysis as the objective of research and as a research method. This covers data especially from Linguistics, Psycholinguistics, Corpus Linguistics, Syntax, Semantics, Lexicography, etc., but also includes other areas such as literary studies, political sciences, and history. An appraisal and selection document is publicly available at ar.sfb833.uni-tuebingen.de/downloads/TALAR_ResearchData.pdf

As part of the Text+ consortium, the repository is part of a network of competence centres with different specializations. As part of the consulting efforts with the users, we establish contact to the best fitting repository within the network. Within Text+, there is always one partner that provides an archive solution for research data from the Text+ community that does not fit any specialization („Catch all“). The fallback partner is SUB Göttingen.

Prior to ingesting a dataset into the repository, quality checks of data are performed by repository staff, members of the General and Computational Linguistics Department, or other local experts in the field. External data is only accepted in the repository if the project seeking deposition of the data has been externally reviewed, for example in a grant application process. In the case where a dataset cannot be reviewed locally, the repository may recommend another CLARIN center with a collection profile which is better matched to the dataset (https://www.clarin-d.net/en/disciplines). Currently, many of the datasets in the repository are widely-used language resources created locally by the Seminar für Sprachwissenschaft (e.g. GermaNet and TüBa-D/Z), or externally created datasets that are well known within the community (e.g. Index Thomisticus Treebank).

Depositors are encouraged (but not forced) to use formats listed in the CLARIN standard recommendations (http://www.clarin.eu/content/standards-and-formats/) when possible. Use of these formats will ensure that the data is interoperable within the CLARIN infrastructure. If possible, data stored in other formats will be converted to an acceptable format before it is archived.

The depositor, with assistance from a data manager if necessary, creates CMDI (http://www.clarin.eu/cmdi/) metadata using components and profiles stored in the Component Registry (http://catalog.clarin.eu/ds/ComponentRegistry/). Metadata is manually checked for completeness and correctness, and automatically validated for syntax and consistency prior to ingestion. Metadata is required, since data without metadata is technically not accepted in the system.

Users of Bagman auto-create CMDI-based metadata by filling-out forms that request information about the research data. When users manually provide metadata, they are asked to instantiate a CMDI-based schema that is supported by the repository and that best matches the type of research data to be archived. All metadata instances are automatically validated against their schema. Only validated metadata is accepted by the repository.

The repository discourages the removal of research data. However, in the case of legal requirements, the archive managers can restrict access to, or even remove the data. When data is removed, the PID originally pointing to it shows a "tombstone" page that explains that the data is no longer available. So far, no removal of research data has been requested.

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Concur No suggestions

Reviewer 1:

Compliance level:

Not Applicable - 0

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:

The core repository infrastructural software is Fedora Commons 4 (FC4), running on a server at the central data center (ZDV) of the University of Tübingen. Fedora Commons software is compliant with the OAIS reference model, and its use promotes adherence to OAIS in all functions of the repository. The ingestion of research data in FC4 invokes documented storage procedures of the underlying software. Fedora has been configured to write all data to dedicated parts of the file system.

The repository does not manage archival copies but the entire content of the repository is regularly submitted to backup procedures. The repository's preservation policy (https://uni-tuebingen.de/en/137029) includes local and distributed backups, reinstalling the repository from backup, and integrity tests of stored data. The processes and procedures of the repository (including ingest, metadata checks, recovery from backup, etc.) are documented in an internal wiki. Locally developed software is stored and documented in an internal git repository. Both the wiki and the git repository are backed up regularly.

The Information, Communication, and Media Center (IKM) of the University of Tübingen is the central information center of the university. It is formed through cooperation between the university library (UB) and the university computing center (ZDV), and reports directly to the rectorate of the University of Tübingen. The computing center of the university provides all central IT-services, including data storage. Storage service is provided in cooperation with the Universities of Stuttgart and Hohenheim under the umbrella of a state internal cooperation plan. The repository makes use of this central infrastructure for backup and operating services.

The repository is physically located at the Central Data Center (ZDV) at the University of Tübingen. They provide the server and its maintenance (including system updates, firewalls, etc), storage, regular hardware checks, and remote backup services. The repository server is configured to perform daily backups to the University of Ulm data center, appoximately 90 km away. Physical access to the servers is allowed only by authorized ZDV staff and is strictly controlled.

In order to maintain the integrity of archived data, checksums based on both the MD5 and SHA256 algorithms are calculated and the stored objects are assessed regularly. In addition, checksums are automatically computed each time a data stream is downloaded. Deviations are visible to the archive managers for taking immediate action.

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

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:

The repository's preservation policy (https://uni-tuebingen.de/en/137029) includes local and distributed backups, the reinstalling of the repository from backup, and integrity tests of stored data. The repository and backups are located in dedicated computing centers with strict access control, and administrator access to the repository is limited to a small group of trained experts.

Depositors are encouraged to use formats listed in the CLARIN standard recommendations (http://www.clarin.eu/content/standards-and-formats/) when possible. The list of accepted data formats may be extended to include new, widely-used formats in the field. In this case, the repository staff will determine which datasets it would be feasible and possible to convert. For example, the repository has converted many of its treebank resources to a new format that has gained popularity in recent years. In the case that a data format is removed from the list of acceptable formats in the future, every effort will be made to convert datasets into an acceptable format.

The Information, Communication, and Media Center (IKM) of the University of Tübingen is the central information center of the university. It is formed through cooperation between the university library (UB) and the university computing center (ZDV), and reports directly to the rectorate of the University of Tübingen. The computing center of the university provides all central IT-services, including data storage. Storage service is provided in cooperation with the universities of Stuttgart and Hohenheim under the umbrella of a statewide concept for data. The repository makes use of this central infrastructure for backup and operating services.

The repository backend was selected for ease of long-term maintenance and compliance to best practice. It has low technical requirements for extracting the resources from the system without additional and proprietary software, making the transfer of the data to new hardware straightforward. Long-term access is ensured by the hardware, open protocols, and organizational embedding in sustainable departmental structures of the university.

Neither the CLARIN-D resource center nor the repository run by it, are legal entities on their own. This also holds for the General and Computational Linguistics Department ("Seminar für Sprachwissenschaft", SfS) where they are located. All are part of the University of Tübingen which is a legal entity - specifically, like all public German universities, a Körperschaft des öffentlichen Rechts, an institution governed under public law.

Depositors must sign a Depositor's agreement (https://uni-tuebingen.de/en/134320) with the University of Tübingen, which ensures that they own all necessary rights required to deposit the data, that they are in compliance with all relevant national and international legal regulations, and that they grant the repository permission to distribute the data in accordance with the access model chosen (public, academic, or individual). Data providers retain all intellectual property rights to their data. In case a violation of conditions is observed, steps will be taken to ensure that the data is not distributed until the issue can be resolved.

The Depositor's agreement (https://uni-tuebingen.de/en/134320) has provisions in place in case research data needs to be migrated to other archives. It also states that the repository has the rights to copy, transform, and store the items, as well as provide access as required. The depositor is also informed that the repository has no responsibility to perform data curation but the repository may do so.

The repository is currently in the process of developing its future strategy, allowing it to guarantee preservation periods to data depositors. Discussions with the BMBF, the state of Baden-Württemberg, and the University of Tübingen are ongoing. An internal working group has been set up to monitor ongoing technical developments of the repository and develop a plan for migrating the repository to newer versions of the repository backend. In this regard, the repository team has documented a potential migration of research data from one repository to another:

- Thorsten Trippel, Claus Zinn: Lessons learned: on the challenges of migrating a research data repository from a research institution to a university library. Language Resources and Evaluation, volume 55, pages 191–207, 2021. See https://doi.org/10.1007/s10579-019-09474-4.

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

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 General and Computational Linguistics Department of the University of Tübingen develops and maintains language resources and tools for community use (https://uni-tuebingen.de/en/134257). Many of these widely used resources have been made available via the Tübingen CLARIN-D repository. These resources, as well as the external resources currently hosted, each have a dedicated webpage containing a detailed description of the resource and contact information for users to ask questions, request a license if required, or give feedback about the resource. Links to the webpage are recorded in the CMDI metadata so that the webpage can also be found using CLARIN search applications.

As part of CLARIN, the repository also participates in additional channels through which members of the designated communities can give feedback on data and metadata hosted by certified centers. These channes include the metadata portal CLARIN Virtual Language Observatory (https://vlo.clarin.eu/), the CLARIN-D Help Desk, and discipline-specific working groups.

The CLARIN Virtual Language Observatory (VLO, https://vlo.clarin.eu/) harvests ISO 24622-1 conformant metadata (CMDI) and displays the large number of available resources through faceted browsing and search facilities. Both in the overview, i.e. when browsing or searching for relevant resources, and on the individual resource pages displaying further information on a specific resource, the user can report an issue or give feedback on metadata records or resources using a designated button connected via a form to the CLARIN-D Help Desk.

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

Furthermore, the discipline-specific working groups (https://www.clarin-d.net/en/disciplines) within the CLARIN-D project are yet another communication channel, through which the various designated communities can provide more general input and feedback on data and metadata to ensure CLARIN-D centres provide relevant resources and resource descriptions. Those discussions will continue under the Text+ umbrella.

The metadata profiles used by the CLARIN-D centre in Tübingen have been selected for descriptive appropriateness for the data types deposited in the repository. ISO 24622-1 provides the framework for selecting these metadata profiles.

To ensure quality of metadata, the repository uses strict schema validation for all provided metadata records. In regular intervals, 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 for instance, our actualisation of CMDI-based profiles from first generation profiles to second-generation profiles.

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 Tübingen repository typically ranks among the top repositories based on a combined score of all evaluated features (https://curation.clarin.eu/collection/table).
Further information about the curation module can be found at https://office.clarin.eu/v/CE-2016-0742-CLARINPLUS-D2_1.pdf.

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, https://vlo.clarin.eu) 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 National 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.

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Clear documentation of the expertise within the repository and the expertise that the repository has access to within the network of partner organizations

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R12 Workflows

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The repository implements explicitly defined workflows described in the repository's preservation policy (https://uni-tuebingen.de/en/137029). Workflows for both depositing data and accessing data are summarized on the repository website (https://uni-tuebingen.de/en/134314). The depositing workflow consists of packaging the resource, creating metadata, and a quality check of data and metadata including PID (Persistent Identifier) assignment. The access workflow includes options for restricting data access. Where possible, processes in the workflow are automated and enforced by the Fedora Commons back-end or through a custom front-end to Fedora Commons developed by the Tübingen CLARIN-D repository staff.

Detailed workflow documentation is maintained in an internal wiki. A general description was also published by Dima, E. et al. (2012): A Repository for the Sustainable Management of Research Data (http://lrec-conf.org/proceedings/lrec2012/pdf/470_Paper.pdf). In brief, the workflow for the ingestion of linguistic resources into the repository system consists of three main phases:
1.) Preparatory Phase: identification of research data; decision on their granularity (unit to be archived) and type (linguistic resources, scientific publication, other data).
2.) Digital Object Creation Phase: hierarchical organisation of research data; association with appropriate organisational unit; selection of resource type, and upload of resources; definition of access restrictions; provision of CMDI-based metadata; preview of all givens and submission to archive management.
3.) Validation and Archiving Phase: checking for correctness and completeness of research data and their metadata description; acceptance or rejection of submission; in case of acceptance: provision of time-stamp and persistent identifier for archival object; archiving of resource. In case of rejection: communication with submitting party to correct/complete submission

Once the research data is archived, its metadata is automatically disseminated via the repository's OAI-PMH protocol. Its metadata becomes locally visible at the repository's website, and globally visible in CLARIN's Virtual Language Observatory.

For each metadata record, a static HTML page is generated upon ingest, including semantic annotation that supports commercial search engines (such as Google) to index the data.

Since the last application for the CTS, a web-based software has been implemented to support researchers in this workflow (https://weblicht.sfs.uni-tuebingen.de/bagman/). The software makes use of the BagIt File Packaging Format (https://datatracker.ietf.org/doc/html/rfc8493). Once a bag of research data and its metadata is submitted via Bagman to the repository, a Python-based script supports archive managers in automatically verifying that the bag has been transferred in a correct and complete manner (the bag contains all files, and all files are non-corrupted). Bagman is described in: Claus Zinn, Bagman – A Tool that Supports Researchers Archiving Their Data, Proceedings of CLARIN Annual Conference Proceedings, Virtual Edition, pages 119-123, 2021 (https://office.clarin.eu/v/CE-2021-1923-CLARIN2021_ConferenceProceedings.pdf).

The repository and its backups are located in dedicated computing centers with strict access control, and administrator access to the repository is limited to a small group of trained experts. This ensures that the data storage and backup is always managed by professionals.

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Clear documentation of workflows, including publicly accessible documentation The balance between publicly accessible information and internal process documentation is reasonable

Reviewer 1:

Compliance level:

Not Applicable - 0

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:

All CLARIN centres (https://www.clarin.eu/content/overview-clarin-centres) provide their metadata in the CMDI format.
The Component MetaData Infrastructure (CMDI) (https://www.clarin.eu/content/component-metadata was initiated by CLARIN to provide a flexible framework for describing metadata based on components and concepts. Each metadata record is based on a profile that is registered in the CLARIN CMDI Component Registry (https://catalog.clarin.eu/ds/ComponentRegistry). Profiles can make use of components. Those building blocks are also registered in the CMDI Component Registry and describe specific aspects or properties of a resource. Elements of CMDI records link to concept definitions that are stored in external registries (like the CLARIN Concept Registry, https://openskos.meertens.knaw.nl/ccr/browser/). Since different communities use different names for the same concepts, linking CMDI elements to concepts enables communities to stick to their terminology while enabling users to find concepts independent of the naming.

A strict requirement for CLARIN centres, and therefore for the Tübingen repository as well, is to make metadata for all resources available through the established and well documented Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (https://www.openarchives.org/pmh/). This standard enables harvesting of the metadata from the repository via http(s).

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

Metadata:
The CLARIN Virtual Language Observatory (VLO) (https://vlo.clarin.eu) of the CLARIN-ERIC harvests the metadata in CMDI format from all CLARIN centres via OAI-PMH. The OAI-PMH endpoint for the Tübingen repository is available at https://talar.sfb833.uni-tuebingen.de/erdora/rest/oai. Due to security concerns, we have limited access to the endpoint to a restricted number of IP addresses. The VLO harvester, for instance, has access to the endpoint. At the time of writing, 563 data sets from TALAR as discoverable through the VLO (see
https://vlo.clarin.eu/?0&fq=collection:T%C3%BCbingen+Archive+of+Language+Resources+%28TALAR%29&fqType=collection:or).

Metadata from CLARIN centres (and other relevant archives and repositories) are browsable and searchable via the VLO website. CLARIN has defined a set of facets to narrow down the selection of resources in the VLO. These facets are again based on concept sets and allow access to potential heterogeneous metadata stocks. The search in the VLO combines a full text query with a selection of (multiple) values in facets.

Data:
For a subset of resources of the CLARIN-infrastructure a “deep search” within the actual data is supported by the means of the CLARIN Federated Content Search (http://www.clarin.eu/fcs) interface. The Tübingen CLARIN-D Repository also offers this kind of access for some of its resources.

PIDs:
The repository uses the common CLARIN PID service (https://www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf) based on the Handle System (http://www.handle.net/) and in cooperation with the European Persistent Identifier Consortium (EPIC). The usage of PIDs is mandatory for resources in CLARIN thus all resources added to the repository may be referenced using PIDs.

CLARIN has a contractual relationship with GWDG concerning the provision of PID-services via EPIC API v2 as mentioned in R0 on repository context.

Citation:
Necessary information for citing resources of the repository - such as PID, resource name, and responsible organization - can be found on the respective page of the VLO, as well as in the metadata. Discussions with CLARIN-ERIC have been initiated on integrating recommended data citations into the VLO.

The Tübingen repository provides preliminary recommendations on how to cite data. This citation recommendation is displayed in the HTML rendering of the metadata. This includes the landing page of the resource, the name of the resource, its creator, release date, and an actionable persistent identifier (in the form of a URL). The landing pages for the General and Computational Linguistics Department resources are also being reviewed and updated to include standard citation recommendations where they are not already listed.

The repository team is working on increasing the findability of the research data it stores. For this, existing metadata profiles and their instances have been enriched with VIAF, GND, ORCID, ISNI identifiers to uniquely refer to persons and organisations involved in the creation of linguistic resources. In addition, all CMDI-based metadata instances are automatically converted into the Linked Open Data Format JSON-LD. As part of the interoperability work, references from CMDI profiles to data categories in the CLARIN concept registry have been replaced by references to schema.org, a widely used light ontology. The repository now complements CMDI-based metadata (directed to the CLARIN VLO) with JSON-LD-based metadata (directed to a much wider community). More details on the conversion can be found in:

- Nino Meisinger, Thorsten Trippel, and Claus Zinn. Increasing CMDI's Semantic Interoperability with schema.org. Accepted for publication in: Proceedings of the Conference on Language Resources and Evaluation (LREC), Marseille, 2022.

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 Tübingen repository is participating in the development of these applications and services and will integrate its data inventory in it.

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R14 Data reuse

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

Tübingen CLARIN-D Repository

All CLARIN centres (https://www.clarin.eu/content/overview-clarin-centres) provide their metadata according to ISO 24622-1 (CMDI) via OAI-PMH. The Component MetaData Infrastructure (CMDI; https://www.clarin.eu/content/component-metadata) was initiated by CLARIN to provide a flexible framework for describing metadata. With this metadata framework it is possible to create metadata schemas tailored to each specific type of resource. It also allows inclusion of all pieces of information deemed useful for potential future users to understand the data in a reuse scenario. This also includes basic information utilized by research data search engines. To avoid proliferation and to allow for transparent structures this is technically realized by bundling descriptive categories into components that can be reused for other data types. Components themselves are then bundled into datatype-specific schemas called profiles. Each metadata record is based on a profile that is registered in the Component Registry (https://catalog.clarin.eu/ds/ComponentRegistry). The data category of CMDI records link to concept definitions that are stored in external registries (like the CLARIN Concept Registry (https://openskos.meertens.knaw.nl/ccr/browser/)). Since different communities use different names for the same concepts, linking CMDI data categories to concepts enables communities to retain their terminology while enabling users to find concepts independent of the naming.

The Tübingen CLARIN-D repository uses the following CMDI profiles, which were designed with concept registry links:

First generation profiles used:

WebLichtWebService: clarin.eu:cr1:p_1320657629644
Resource Bundle: clarin.eu:cr1:p_1320657629649
OLAC-DcmiTerms: clarin.eu:cr1:p_1288172614026
DcmiTerms: clarin.eu:cr1:p_1288172614023

Second Generation Profiles:

ExperimentProfile:clarin.eu:cr1:p_1447674760337
TextCorpusProfile: clarin.eu:cr1:p_1442920133046
LexicalResourceProfile: clarin.eu:cr1:p_1445542587893
SpeechCorpusProfile: clarin.eu:cr1:p_1485173990943
CourseProfile: clarin.eu:cr1:p_1505397653792
ToolProfile: clarin.eu:cr1:p_1447674760338

The second generation profiles extend the first generation profiles, for example by allowing metadata files to include authoritative IDs for individuals and institutions, such as VIAF links and ORCID IDs. The repository is currently updating metadata files from first generation profiles to second generations profiles.

Prior to ingest, the depositor and the data manager check that all relevant metadata fields have been filled in correctly, and as completely as possible. Special attention is given to these components:

- General Information, including name of the resource, type of resource, version of the data, life cycle status,legal owner, start year, field of research, modality
- Project, including relevant information about the project in which the resource was created, if it was created within a specific research project
- Publications, which describe the resource or are based on the resource
- Creation, including the information on each individual involved in the resource creation, software tools used in the creation of the dataset, and third party data contained in the dataset
- Documentation of the resource, i.e. external descriptions of the resource
- Access Information, i.e. under which licence a data user may receive the resource, associated software that can be used to work with the data
- Technical Information for each file part of the data set
- Resource Type Specific Information, such as the size of a text collection in terms of number of words or the number of recording hours for speech corpora

The components for General Information, Project, Creation, Access and Technical Information are required in most profiles, and the data manager ensures that those are completed. Data Type Specific Information is strongly encouraged by the data manager. Since Documentation, Project, and Publications may not be available for each resource, they are optional.

Depositors are encouraged to use formats listed in the CLARIN standard recommendations (http://www.clarin.eu/content/standards-and-formats). Use of these formats will ensure that the data is interoperable within the CLARIN infrastructure. If possible, data stored in other formats will be converted to an acceptable format before archiving. In the case that a particular format is replaced by a more widely-used format, data will be converted and archived under a new PID. The landing pages for resources developed at the General and Computational Linguistics Department of the University of Tübingen contain a detailed description of the resource, including which data format(s) are used and any associated software that can be used with the data (e.g. TüBa-D/Z treebank: https://uni-tuebingen.de/en/134290; GermaNet wordnet: http://www.sfs.uni-tuebingen.de/GermaNet/index.shtml). Resources have been, and will continue to be, converted to new formats as the need arises.

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Concur Clear and complete documentation

Reviewer 1:

Compliance level:

Not Applicable - 0

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:

As part of CLARIN-D we are committed to play an active role in the development of CLARIN's repository infrastructure. General plans for maintaining and further developing the infrastructure have been formulated as part of the project proposal. This active role will continue within the Text+ project.

The central goal is to improve the usability of the research infrastructure for typical research tasks such as the retrieval of resources, the evaluation of data or the publication of results. To achieve this, modifications and extensions to a variety of infrastructure components in the repository and in the central infrastructure are necessary. Meetings of all centres to monitor advances in infrastructure development take place quarterly.

Further important goals of infrastructure development (https://www.clarin.eu/content/clarin-technology-introduction) are:
- To ensure resilience, integrity, and availability of the sustainable repositories and the central infrastructure
- To integrate new resources and tools based on the needs of the user communities
- To allow for better interoperability of tools and resources in the infrastructure
- To enhance the central content search to be more useful in actual research tasks
- To optimize metadata of the resources provided and to enhance user experience in central metadata search

Additional strategic infrastructure planning takes place on the European level in the coordinating committee of the technical centres of the CLARIN ERIC where CLARIN-D also participates.

The repository adheres to all standards and best practice recommendations set forth by CLARIN, as well as meeting the requirements of OAIS, as described in the preservation policy (https://uni-tuebingen.de/en/137029).

The repository is built on a reliable and stable technical infrastructure, which has been tested and evaluated and has been determined to be fully functional for the needs of the repository. The main technical components used by the repository include:

- The core repository infrastructural software is Fedora Commons 4, running on a server at the central data center (ZDV) of the University of Tübingen. Fedora Commons software is compliant with the OAIS reference model, and its use promotes adherence to OAIS in all functions of the repository.

- At the time of writing, the server (itself a Virtual Machine) runs the operating system AlmaLinux release 8.5 (Arctic Sphynx), see https://almalinux.org, and is updated as needed.

- Server maintenence and geographically distributed daily backups (https://uni-tuebingen.de/en/2944) are performed by the ZDV.

- The repository has developed and maintains a customized administrator interface to the Fedora Commons backend that aids in carrying out the workflows and functions described in the preservation policy.

- All locally developed software is housed in a local git repository, and all workflow documentation in an internal wiki. Both the git repository and the wiki are backed up regularly by the local system administrator.

- Firewalls permit only authorized access to the systems on which the repository is operated, including access to administrative tools and backends from authorized workstations.

- Icinga (https://www.icinga.com/) is used by CLARIN-D (http://clarin-d.net/images/ap3/ap3-005-monitoring.pdf) to monitor infrastructure components, including repository probes. Repository probes are made every few minutes and all repository technical staff are notified by email if a problem occurs so that it can be resolved quickly.

- Metadata is created according to ISO 24622-1 and ISO 24622-2 standards, using either the web based tool Bagman (https://weblicht.sfs.uni-tuebingen.de/bagman/) or standard XML editors such as Oxygen. A set of XQuery functions are used to test and generate reports on the quality of the CMDI metadata. These tests include consistency checks (use of PIDs, naming of persons and institutions, field names, etc), file record reference update notifications, completeness, etc. Research data submitted via Bagman are automatically tested for completeness and correctness.

- Metadata is disseminated using the OAI-PMH protocol, with the PROAI plugin of Fedora Commons.

- PIDs are aquired from, and resolved by, the Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), with whom CLARIN has a contractual relationship.

- The repository bandwidth is at least 100 MB/sec, the ZDV guarantees a high availability/uptime for the server and its connection to the Internet.

Data management: The standard Fedora Commons tools, in combination with a custom administration application are used for data management. Metadata is distributed via the OAI-PMH protocol, supporting selective harvesting as well. Both the OAI-PMH supplied metadata and the Fedora Commons tools are used to report on the status of the data.

Administration: Using a local Authentication, Authorization and Access Infrastructure in Fedora Commons, data managers conduct administrative tasks. The hardware is securely stored in locations with highly restricted access.

Preservation Planning: A technology audit is used to evaluate the state of technology, long time efficiency and test migration procedures when new platforms become available. The migration tests are conducted routinely to different hardware even if the productive environment is not migrating. The cooperation with partner projects supports the preservation activities. The open format used by Fedora Commons guarantees the long-term accessibility of the data.

Access: The digital objects are available for reading access via their PID for authorized users, based on the AAI infrastructure of the CLARIN Service Provider Federation and a local user management. The PIDs are available in the metadata, which can be harvested via OAI-PMH (e.g. by the VLO).

The technical infrastructure and processes described here have been tested and evaluated and have been determined to be fully functional for the needs of our repository. State of the art firewalls block unauthorized access to the systems on which the repositories are being operated, including access to administrative tools and backends from unauthorized workstations.

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 (https://www.clarin.eu/governance/standards-committee). 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.

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R16 Security

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The Information, Communication, and Media Center (IKM) of the University of Tübingen is the central information center of the university. It is formed through cooperation between the university library (UB) and the university computing center (ZDV), and reports directly to the rectorate of the University of Tübingen. The computing center of the university provides all central IT-services, including data storage. Storage service is provided in cooperation with the Universities of Stuttgart and Hohenheim under the umbrella of a statewide concept for data. The repository makes use of this central infrastructure for backup and operating services.

The repository's preservation policy (https://uni-tuebingen.de/en/137029) includes local and distributed backups, reinstalling the repository from backup and integrity tests of stored data. The repository and backups are physically located in dedicated computing centers (ZDV/Tübingen, ULM data centre, see below), both with strict access control, and administrator access to the repository is limited to a small group of trained experts. Physical security is hence guaranteed by the aforementioned computing centers, which have in place a number of (potentially non-disclosed) security measures.

The repository runs on a server hosted, managed and maintained by the ZDV, who is also responsible for making daily backups of the data and system configurations to a remote location. The repository server is currently configured to perform daily backups to the University of Ulm data center using Bacula (https://uni-tuebingen.de/en/2944), and a detailed report is sent to the repository staff for each backup.

In case of disaster, recovery will first be attempted through the ZDV backups, and then through the documented recovery procedures of the alternative backup strategy.
We have repeatedly tested the proper functioning of the recovery functionality. In case of data loss, the repository can fully recover within 24 hours. Within this short time frame, the Tübingen CLARIN-D Repository has thus the ability to ensure the continuity of service in case of unexpected events (including potentially malicious attacks aiming at corrupting all data).

Data resilience (and subsequent recovery) is ensured by automated monitoring.
The repository status and availability of resources are continually monitored within the CLARIN infrastructure. In case of any failure, the repository staff is notified immediately, so that any data breaches or other service disruptions can be dealt with in a speedy manner.

In order to maintain the integrity of archived data, checksums based on the MD5 and SHA256 algorithms are being calculated and the stored objects are assessed regularly. In addition, checksums are automatically computed each time a data stream is downloaded. Deviations are visible to the archive managers for taking immediate action.

In sum, the repository provides a high-level of information security for all the data it hosts.

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Applicant Feedback

R17 Applicant Feedback

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

Thank you for reviewing our application.

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:

My overall assessment is that this is a very well-written response from a repository which has clearly fully implemented the guidelines, both in terms of the specifics of those guidelines and the intent which those guidelines represent.