Assessment Information

CoreTrustSeal Requirements 2020–2023

Repository:	CLARIN Center BBAW
Website:	https://clarin.bbaw.de/en/
Certification period:	14 September 2023 - 13 September 2026
Requirements version:	CoreTrustSeal Requirements 2020-2022
This repository is owned by:	**Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)**

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
- Institutional repository

Links:

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

ok

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 Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) [1] has a long-standing tradition of corpus-based lexicography and is committed to open access to its primary research data. The institutional subject-based data repository [2] is a member of the European CLARIN research infrastructure [3] and the research infrastructure consortium Text+ [4] that is a member of the German initiative to establish a national research data infrastructure (Nationale Forschungsdateninfrastruktur – NFDI) [5]. It publishes and preserves historical and contemporary German text corpora as well as the lexical resources provided by the Zentrum Sprache (Language Centre) [6] at the BBAW - mostly in open formats like XML.

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.

CLARIN is committed to boosting humanities research in a multicultural and multilingual Europe, by facilitating access to language resources and technology for researchers and scholars across a wide spectrum of domains in the humanities and social sciences (Krauwer, 2008).

The 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 "Collections", "Lexical resources" and "Editions".

[1] https://www.bbaw.de/en/
[2] https://clarin.bbaw.de/en/
[3] https://www.clarin.eu/
[4] https://www.text-plus.org/en/
[5] https://www.nfdi.de/?lang=en
[6] https://www.bbaw.de/forschung/zentren/zentrum-sprache

Links:

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

ok

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 designated community consists predominantly of scientific researchers (partly historians, lexicographers, psychologists, literary scholars and germanists/specialists in German studies).

Links:

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

ok

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

Links:

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

ok

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:

The level of curation performed depends on the contracts signed (level B-D). Most of the data in the repository is produced by our partner project Deutsches Textarchiv (DTA) [1], these datasets are curated on data-level (level D) by their team at the BBAW CLARIN center [2].
[1] https://www.deutschestextarchiv.de/
[2] https://clarin.bbaw.de/en/

Links:

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

ok

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Insource/Outsource Partners

If applicable, please list them.

Compliance level:

Not Applicable - 0

Response:

List of outsource partners:

1) Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) [1]

The repository makes use of a common CLARIN PID service [2] based on the Handle System [3] 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 [4].

GWDG has been sucessfully certified in May 2013 according to ISO 9001:2008.

2) The repository is 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 [5].

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

* CLARIN-D HelpDesk [6]: 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 [7]: A starting point for researchers to find information on CLARIN-D and to access CLARIN-D services.
* CLARIN-D wiki [8]: A central platform for CLARIN-D-related stuff.
* CLARIN central monitoring [9]: 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) in the future. This includes the helpdesk, monitoring service, and a new documentation platform.

3) CLARIN-ERIC

CLARIN-D is a member of CLARIN'S European Research Infrastructure Consortium (ERIC). CLARIN-ERIC offers central services to it's members and users, as stated here [10].

The services are available to all centres in the member countries of the CLARIN-ERIC [11].

The most important services of the ERIC cover the search functionality for the German CLARIN centres:

* Virtual Language Obervatory - VLO [12]: 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 [13]: Optionally, centres can provide the actual data of their resources for this central content search.
* CMDI Component Registry [14]: CLARIN's registry for components and profiles according to ISO-24622-1.

In addition, CLARIN-ERIC offers several further services such as central registries, user statistics management and, as an official EUDAT community, access to advanced EUDAT services.

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.

[1] https://info.gwdg.de/dokuwiki/doku.php?id=en:services:it_consulting:scientific_data_management:start
[2] https://www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf
[3] http://www.handle.net/
[4] http://www.clarin-d.de/mwiki/images/0/0b/GWDG_PID.pdf
[5] https://www.clarin-d.net/de/ueber/zentren/zusammenarbeit
[6] https://support.clarin-d.de/mail/
[7] https://clarin-d.de/en/
[8] https://www.clarin-d.de/mwiki/index.php/Hauptseite
[9] https://monitoring.clarin.eu/
[10] https://www.clarin.eu/value-proposition
[11] https://www.clarin.eu/content/overview-clarin-centres
[12] https://vlo.clarin.eu
[13] https://www.clarin.eu/contentsearch
[14] https://catalog.clarin.eu/ds/ComponentRegistry

Links:

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

ok

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:

-

Links:

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:

ok

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, and are fulfilled by this resource center:

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

An overview of all requirements for centres of type B is also given in the form of a checklist [1].

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).

All CLARIN centres get a record in the Registry of Research Data Repositories (re3data.org) [2].

[1] https://office.clarin.eu/v/CE-2013-0095-B_checklist-v7_3_1.pdf
[2] https://www.re3data.org/repository/r3d100012054

Links:

Reviews

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

ok

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 this repository is to ensure the availability and long-term data preservation of German historical and contemporary text corpora, and lexical resources provided by the Zentrum Sprache (Language Centre) at the Berlin-Brandenburg Academy of Science and Humanities (BBAW) [1]. It may also serve as a depositing solution for data created by projects external to the BBAW as long as they are freely licensed (e.g. under a Creative Commons type license) and fit well into the portfolio of BBAW research interest.

This mission is supported by the infrastructure of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), by the integration of the repository into the national and international CLARIN infrastructures [2] and has been officially enacted by the representative (president) of the BBAW [3].

[1] https://www.bbaw.de/en/
[2] https://www.clarin.eu/
[3] https://clarin.bbaw.de/en/mission

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:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

ok

R2 Licenses

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The repository is no legal entity in its own right. It is run by the Berlin-Brandenburg Academy of Sciences and Humanities which is an institution governed by public law. Deposits are handled in a case-by-case approach. There are individual contracts and different licenses for each resource we have archived. The access to the items is also handled case-by-case, ranging from open access over restricted access requiring a contract to restricted access on-site. The depositors themselves are responsible for compliance with any legal regulations in the area where the data is collected. Where required by national regulations, the archive also signs contracts with national/regional institutions.

Before ingest and signing the contracts, our staff makes a plausibility check for the data and metadata.

Our contracts cover the following items:

* update/maintenance procedures
* ownership, IPR and liability for it (including personal data see section 4c)
* license types
* compensation/payments
* liability for damages and costs
* termination of the agreements

An example contract can be downloaded here: [1]

According to the contract, the depositor may choose between three access levels: 'unrestricted public', 'academic access only' or 'restricted access/only after receiving permission from the depositor'.

Most of the data in the repository have Creative Commons licenses [2] applied to them (unrestricted access).

We're currently not providing access to personal data. Any personal data will be anonymized during the curation procedure (see also R4).

If the data consumer does not comply with the access regulations, the only measure that can be taken in practice is to deny him/her further access and to make the research community aware of the misuse. For some data sets, explicit permission from the depositor is needed. In that case a login is necessary.

For contracts which require 'academic access only', we rely on the CLARIN AAI identity federation (single sign-on) [3].

For 'restricted access/only after receiving permission from the depositor', we rely on .htaccess rules [4] of the webserver.

We specifically rely on the DFG ethical Codes of Conduct (e.g. layed down in the DFG Rules of Good Scientific Practice):

* ALLEA (ALL European Academies) European Science Foundation, The European Code of Conduct for Research Integrity [5]

* DFG, Rules of Good Scientific Practice [6]

* BBAW, Richtlinien zur Sicherung guter wissenschaftlicher Praxis [7]

Data users have to follow the repository's General Terms of Use [8] which are linked in the data catalog at the bottom of each bibliographic data page and also in the repository description page [9].

[1] https://clarin.bbaw.de/bbaw/static/pub/CLARIN_Template_Depositors_Agreement_BBAW.pdf
[2] https://creativecommons.org
[3] https://www.clarin.eu/content/federated-identity
[4] https://en.wikipedia.org/wiki/.htaccess
[5] https://allea.org/code-of-conduct/
[6] https://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice/
[7] https://www.bbaw.de/files-bbaw/die-akademie/dokumente/AKTUELL_Richtlinen_gute_wissenschaftliche_Praxis_2002-06-27.pdf
[8] https://clarin.bbaw.de/bbaw/static/pub/CLARIN_Template_Terms_of_Use_BBAW.pdf
[9] https://clarin.bbaw.de/en/repo/

Links:

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:

ok

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 repository is in the implementation phase - 3

Response:

As stated in the Depositor's Agreement [1] , 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.

CLARIN centres commit to ensuring long-term availability, access and to preservation of datasets submitted to their repositories, as set out in their Mission statements. CLARIN centres are setup as a distributed network, where each centre institution is a hub of the digital humanities and brings its own financial resources into CLARIN, which ensures continued availability. Thus, in case of a withdrawal of funding, the repositories content would be transferred to another CLARIN centre. The legal aspects of the process of relocating data to another institution is addressed by the Depositor's Agreement [1].

The CLARIN-D centres have signed a memorandum of understanding [2] to confirm that all CLARIN-D centers are willing to take over each others repository contents in case a center ceases to exist. Also the cooperation agreement of the CLARIN-D centers is available online [3].

The BBAW 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. The general plan of the NFDI is to be the long-term financed cornerstone of Germany's future research infrastructure. [4]

Furthermore, the BBAW is member of the "Geistes und kulturwissenschaftliche Forschungsinfrastrukturen e.V." [5], 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.

For the BBAW CLARIN centre and its repository, funding has been secured until at least 09/2024 in the context of the Academies's Programme (funded by the umbrella organization of the German academies). From 09/2024 onwards, the repository will be taken over by the NFDI initiative where BBAW is a partner of the TEXT+ consortium [6] which is currently funded until 09/2026. The consortium builds on expertise and services from both CLARIN and DARIAH and aims to extend into NFDI data centres in the long-term. A certified repository in turn is a required building block for such data centres. Thus, the repository's funding is secured until at least 2026 and its continuity assured by being a cornerstone of the BBAW's long-term plans in the context of the NFDI.

With the backing of the repository by 2 funding sources and the transition of the repository from the CLARIN into the NFDI/Text+ context, the appointed work force sums up to 2,25 FTE positions financed until 2026.

[1] https://clarin.bbaw.de/bbaw/static/pub/CLARIN_Template_Depositors_Agreement_BBAW.pdf
[2] https://www.clarin-d.net/en/about/centres/mou-taking-other-centre-s-data
[3] https://www.clarin-d.net/en/about/centres/division-of-labour-of-clarin-d-centres
[4] https://www.text-plus.org/en/
[5] http://www.textgrid-verein.de/
[6] https://www.bbaw.de/forschung/text-plus

Links:

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Level 3 because of the uncertainty of funding and not having a published continuity plan in place.

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

There's a good succession agreement in place with the MoU.

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 repository includes resources provided by CLARIN-D member institutions and other institutions and/or organizations that belong to the CLARIN-D extended community. The data in our repository contains sufficient information for others to assess the scientific and scholarly quality of the research data in compliance with disciplinary and ethical norms. We specifically rely on DFG ethical Codes of Conduct (e.g. layed down in the DFG Rules of Good Scientific Practice). Our repository does not (and cannot) systematically verify whether the data received have been collected according to these quality standards, but the depositor needs to state it in the depositors contract.

Disclosure risk is minimized by anonymization or limited access via login accounts.

According to the contract, anonymization of the datasets must be done by the depositor and 'the Depositor guarantees that Content contains no data or other elements that are contrary to the law or public regulations' [1].

In case during the archiving workflow (in the diagram [2] at 'DTA inspection and feedback') our staff would find data with disclosure risk, the data would either be rejected until anonymized by the depositor or it would be saved with limited access via login accounts according to the contract. Our staff would provide guidance to how anonymization would be done properly. There is an internal checklist available on anonymization and actions taken in previous cases, based on recommendations from the DFG Handout "Information on legal aspects when dealing with language corpora" (german only) [3].

We specifically rely on the DFG ethical Codes of Conduct (e.g. layed down in the DFG Rules of Good Scientific Practice):

ALLEA (ALL European Academies) European Science Foundation, The European Code of Conduct for Research Integrity [4]

DFG, Rules of Good Scientific Practice [5]

BBAW, Richtlinien zur Sicherung guter wissenschaftlicher Praxis [6]

[1] https://clarin.bbaw.de/bbaw/static/pub/CLARIN_Template_Depositors_Agreement_BBAW.pdf
[2] https://clarin.bbaw.de/en/repo/#workflow
[3] https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_recht.pdf
[4] https://allea.org/code-of-conduct/
[5] https://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice/
[6] https://www.bbaw.de/files-bbaw/die-akademie/dokumente/AKTUELL_Richtlinen_gute_wissenschaftliche_Praxis_2002-06-27.pdf

Links:

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:

ok

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text:

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 Berlin-Brandenburg academy of sciences and humanities (founded in 1700) is a recognized institution for longterm projects. The BBAW and its repository is a member of the Text+ consortium as a part of Germany's National Research Data Infrastructure (NFDI) [2] 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.

In addition our repository is part of CLARIN [3], a research infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences. CLARIN also offers information on a wide range of topics, including teaching material, help on data management plans and other, discipline-specific support.

As stated in R3, for the BBAW CLARIN service center as a whole [1] funding is secured until 09/2024 , for its repository and associated staff, funding is secured until at least 2026.

The repository staff consists of scientists with solid knowledge of and experience in the field of the digital humanities data management. They are organized in three functional groups: Administration, Technology, and Data curation. The staff consists of part-time appointees such that the work force sums up to 2,25 FTE positions financed by BBAW via the Academies' Programme and its NFDI initiative.

Administration
Project leader, Board reporting (Computational Linguist, 0,5 FTE)

Technology
Software Developer, web administration (, Research Software Engineer, 0,25 FTE)
Systems Administrator, Systems, networking, hardware and software, ingest of data (IT specialist, 0,25 FTE)

Data curation
Data Managers (2), Data provider relations, data conversion, quality checks, protection of respondent privacy (Philologist, Computational Linguist 1,25 FTE)

The repository staff members have access to training on data management, metadata, long-term preservation and professional development (offered by CLARIN and Text+, see [4], [5], [6]). This includes budget for regular developer meetings, mobility grants for sharing of expertise, conferences, workshops, meetings with their respective scientific communities as well as a centralized knowledge base (user guide [7], wiki, bugtracker and mailing lists).

[1] https://clarin.bbaw.de/en
[2] https://www.nfdi.de/?lang=en
[3] https://www.clarin.eu/
[4] https://www.clarin.eu -> “Learn & Exchange”
[5] https://www.text-plus.org/en/research-data/data-and-competence-centres/
[6] https://www.text-plus.org/en/networking/cross-cutting-topics-2/
[7] https://www.clarin-d.net/en/language-resources-and-services/user-guide

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Accept with note that the repository has adequate funding until 2026.

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

The additional information provided in R3 and R5 helped to bunderstand BBAW's funding situation much better, the reviewer sees no need for the warning about FTE decrease anymore.

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 [1].

Text+ is structured along so-called “data domains” which are organized in thematic clusters.The repository is part of the data domains "Collections", "Editions" and “Lexical resources”. For each of those exists a Scientific 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.

We participate in the CLARIN/Text+ help desk [3] 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.

The BBAW also has multiple advisors:

1) The Berlin-Brandenburg Academy of Sciences and Humanities elects its currently 380 members from all over Germany and from abroad from all disciplines. Anyone who has distinguished himself through scientific achievements can be appointed as a member. These members (partly historians, lexicographers, psychologists, literary scholars and germanists/specialists in German studies) can be considered as in-house experts and are asked for relevant text sources in the design phase of reference corpora which are deposited in the repository. For these groups our text corpora and lexical resources are particularly relevant. The current members can be seen in this list [4].

2) There are multiple ways how BBAW receives feedback from it's designated community:

* more than 2000 active users work in the DTAQ collaborative web curator tool, description [5], login [6]
* via the issue tracker on Github for the DTA base format (DTABf) [7]
* via the DTA Twitter channel [8]
* we get feedback on publications [9]
* our cooperation partners also give feedback [10]
* in journal articles [11]
* we get feedback on conferences and workshops

[1] https://www.clarin.eu -> “Learn & Exchange”
[2] https://www.text-plus.org/en/about-us/coordination-committees-2/
[3] https://support.clarin-d.de/otrs/
[4] https://www.bbaw.de/en/the-academy/members
[5] https://www.deutschestextarchiv.de/misc/2013-04_poster_allea/poster.pdf
[6] https://www.deutschestextarchiv.de/dtaq
[7] https://github.com/deutschestextarchiv/dtabf/issues
[8] https://twitter.com/textarchiv
[9] https://www.deutschestextarchiv.de/doku/publikationen
[10] https://deutschestextarchiv.de/clarin-kooperationen
[11] https://ride.i-d-e.de/issues/issue-6/deutsches-textarchiv/

Links:

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:

ok

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Excellent expert guidance

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:

To ensure non-corruption of the data, data is always validated by XML tools before ingestion. The integrity of the data is ensured by the version control mechanism in the Fedora-Commons [1] back-end by MD5 checksums. Checksum tests are done regularly, especially before performing backups.

Our software workflow allows to ingest data only if metadata (in CMDI [2] and DC [3] format) also is present.

Most of the data in the repository is ingested by the Deutsches Textarchiv (DTA [4]) which uses version control mechanisms (provided by GIT [5] version control software) to log changes. The repository itself currently does not log differences in metadata or data. Completeness of metadata is verified during the DTA workflow [6].

Data and metadata in the repository are considered as fixed and immutable. New digital objects and persistent identifiers are created for updates. All previous versions of a newly submitted existing digital object can be reached via links in the dropdown object history in the web frontend.

According to the contract [7], 'the repository has the right to modify the format and/or functionality of Content if this is necessary in order to facilitate the digital sustainability, distribution or re-use of content.' Generally new versions are only ingested if content differs from previous versions.

Provenance is documented through the CMDI [2] metadata description page.

All archived objects are linked to their metadata descriptions and are organized in tree structures to indicate relationships between objects.

Essential properties of different versions of the same file are checked via checksums. In case there is no difference, updates will not be ingested.

The identities of the depositors are checked by the repository staff when they hand over their data.

Some examples for versioning, object history and audit trails of updated data sets can be seen here:
DTAQ Versioning Example [8]
Fedora Object History Example [9]
Fedora Audit Trail Example [10]

[1] https://fedorarepository.org/
[2] https://www.clarin.eu/cmdi
[3] https://dublincore.org/
[4] https://www.deutsches-textarchiv.de/
[5] https://git-scm.com/
[6] https://clarin.bbaw.de/bbaw/static/img/Archiv_Workflow_en.jpeg
[7] https://clarin.bbaw.de/bbaw/static/pub/CLARIN_Template_Depositors_Agreement_BBAW.pdf
[8] https://clarin.bbaw.de/bbaw/static/pub/Ursel.pdf
[9] https://clarin.bbaw.de:8088/fedora/objects/dwds:1/versions
[10] https://clarin.bbaw.de/bbaw/static/pub/objectXML.pdf

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

ok

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Integrity checks conforms to state of the art processes, good pre-ingest checks are in place
For next application, you might consider to provide more detailed information, eg screen shots, about the object history, audit trail and event metadata.

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:

A collection development/appraisal document describing the scope of data curation at the BBAW CLARIN center repository is available here [1].
Generally the repository accepts historical and contemporary German text corpora as well as the lexical resources provided by the Zentrum Sprache (Language Centre [2]), but the Deposits are handled in a case-by-case approach.
Quality checks are done before ingest by our staff. See the workflow image [3].
Without proper metadata, the software allows no ingest of data. Therefore the presence of metadata is also checked by our staff before ingest.
A minimum set of TEI metadata which we generate DC and CMDI metadata from can be found here [4].
The repository provides a list of accepted formats, including common multimedia-document formats as well as formats for binaries. For other file formats, we provide advice for conversion.
CLARIN recommendations for data-deposition formats [5]
CLARIN centers Standards Information System [6]
For data which does not fall within the collection profile the repository team would recommend another CLARIN center with a collection profile which is closer to its discipline [7] to the depositor.
The number of accepted file formats is limited, to make future conversions to other formats more feasible. Open (non-proprietary) file formats are used whenever possible. For textual resources, XML formats are used whenever possible, to ensure future interpretability of the files independent of the tool used to create them. Text is encoded in Unicode to ensure future interpretability.
According to the repository documentation in the workflow image ([3] on the right hand side), our staff inspects the depositor's files to ensure that the files meet the repositories requirements. If this is not the case, then our staff gives feedback and allows the depositor to generate valid files and metadata.
In a rare case where an item has to be removed from the collection we would keep its metadata, point its persistent identifiers to it to keep them functional and notify the depositor.
[1] https://clarin.bbaw.de/en/curation/#Contents
[2] https://www.bbaw.de/forschung/zentren/sprache
[3] https://clarin.bbaw.de/bbaw/static/img/Archiv_Workflow_en.jpeg
[4] https://hdl.handle.net/21.11120/0000-0000-F4B5-0
[5] https://www.clarin.eu/content/standards
[6] https://standards.clarin.eu/sis/
[7] https://www.clarin-d.net/en/disciplines

Links:

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:

The guideline has been fully implemented in the repository - 4

Comments:

ok

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:

BBAW CLARIN Centre preservation policy can be found here [1].

An internal wiki with description and manuals for usage of storage and backups systems is maintained. Only trained admins have access to the systems. Also to coordinate admin activities (e.g. to restore possibly corrupted files) a ticketing system is maintained to document each step taken. A complete change history of managed content is maintained this way as it may evolve over time. This ensures the procedures are carried out in the required way.

From the website [2]:

"The virtual machines used by the BBAW CLARIN center repository reside on hard disk space secured by RAID level 6. Every night filesystem and database dumps of the virtual machines are copied to a dedicated backup server system (also RAID level 6).

Deterioration of disk media is checked by S.M.A.R.T. status checks. Weekly backups to a LTO-8 tape library are performed. Backup tapes are deposited in a locked safe in a separate fire safety zone of the building. Each year one additional full backup tape set is separated and added to a long term archive.

The backup software (Bareos) internally makes use of checksums to recognize tape block errors [3] and tape media deterioration.

The virtual machines disk images are dumped and replicated to a secondary virtualization server in a different server room in a different fire safety zone. In case of a system failure, these replicated disk images can be manually started within minutes."

Data files for which the BBAW CLARIN center has assumed ownership via contract are periodically reviewed with respect to current applicable file formats. All data files are subject to periodic consistency and validation checks.

Risk management strategies are formulated at the institutional level (in BBAW IT security concept) and include data disaster recovery procedures. See also R16 for further details on risk management.

Consistency checks between backups typically rely on checksums which are calculated on individual metadata and data files.

The integrity of the data is ensured by the version control mechanism in the Fedora-Commons backend (based on internal MD5 hash values). All datastreams and versions are equipped with a MD5 checksum, which is checked in coordination with the backups as described above. Metadata is also a data stream within the digital object, and as such is version controlled like object data.

An overview of the storage locations and media used in the workflow [5] can be found here: [6]

[1] https://clarin.bbaw.de/en/preservation
[2] https://clarin.bbaw.de/en/preservation/#Storage_Procedures
[3] https://docs.bareos.org/Configuration/StorageDaemon.html#config-Sd_Device_BlockChecksum
[4] https://clarin.bbaw.de/bbaw/static/img/Archiv_Workflow_en.jpeg
[5] https://clarin.bbaw.de/bbaw/static/img/storage_locations.pdf

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

ok

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:

We indicate to Data Producers and to our community (via our preservation policy [1]) that we undertake to preserve datasets posted to the repository for the long-term. For this, we rely on our institution (Berlin-Brandenburg Academy of Sciences and Humanities) and on our membership in CLARIN and Text+.

The preservation level is defined as: to ingest/store the data, to provide endorsement for integrity of the data and to ensure the data is accessible and usable to data consumers.

See also R12:
"Archival Storage. This entity is responsible for the systematic storage, maintenance and retrieval of the data. It further performs routine checks on media quality (refresh if necessary), errors and disaster recovery capabilities. Implications for the BBAW CLARIN Center: Two separate functions were implemented: Data management (which is responsible for storage of the data, error detection and retrieval) and system management (which is responsible for media quality and recoverability).
...
Preservation Planning. This entity is responsible for evaluation of quality of service, state of development intechnology and provides migration planning. Implications for the BBAW CLARIN Center: The BBAW participates in digital infrastructure projects like CLARIN and Text+ to monitor the technical developments and also community feedback to provide reliable and useful services. The BBAW service center is continuously updated according to the needs of these projects."

The example depositor's contract [2] provides for all actions to meet the responsibilities (licensing, copying, storing, modification, distribution and custody transfer).

The repository is granted a non-exclusive license of the data by the depositor.

See also R14:
"In 2017 our repository made a transition from CMDI metadata version 1.0 to 1.2. This has been achieved by a standard XSL transformation of previous CMDI metadata versions. Future migrations are expected to be handled in a similar way. Measures are taken to ensure the future interpretability of the data. The number of accepted file formats is limited, to make future conversions to other formats more feasible. Open (non-proprietary) file formats are used whenever possible. For textual resources, XML formats are used whenever possible, to ensure future interpretability of the files independent of the tool used to create them. Text is encoded in Unicode to ensure future interpretability."

The information basis on which the decision is made to decide a preservation action is nescessary includes
- severe errors in transcription or metadata were corrected (according to the - DTAQ collaborative web curator tool)

- CMDI oder TEI XML format update
- technological developments
- community best practice

Our preservation plan is internal and lists several measures to ensure the future interpretability of the data and avoid data loss as a result of risk management.
Scenarios that might put data at risk are listed:
- outdated file format (versions)
- data doesn't conform to format specification (e.g. XML is not validating)
- bitstream corruption
- errors in data or metadata

As countermeasures we would identify the problem and setup a test case to fix a - copy of the data set by eventually:
- choosing a proper conversion tool
- updating file to newer format
- fixing the file structure to make it validate
- restoring the file from backup, check media and compare files
- fixing errors in data or metadata in the DTAQ collaborative web curator tool

If that test setup fixed the problem sucessfully, it is then applied to a copy of all affected datasets and the modified data sets are checked for XML validity and then ingested into an internal test repository. If all went well (the datasets in the test repository passed all of our quality checks), then the modified data sets are ingested in the production repository (keeping the all previous data set versions).

The preservation strategy is to continuously refresh our physical storage media and migrate the data files via conversion to newer standards.

To evaluate tools and their outcome we rely on WebLicht [3] to compare linguistic toolchains. For XML format validation, we rely on Oxygen XML Editor and libxml software library. There is also a paper on analysis and correction of annotation and transcription errors, see [4].

Significant properties are defined and checked via the validity of the XML file structures, for web services we have tests in place to compare their outcome.

Preservation action is documented in the DTAQ collaborative web curator tool as well as via comment entries for the datastreams during ingest, see
DTAQ Versioning Example [5]
Fedora Audit Trail Example [6]
see also
Fedora Wiki: Content Versioning [7]

Existing AIPs stay unmodified in the repository, modified AIPs will be added, keeping all previous data sets.

[1] https://clarin.bbaw.de/en/preservation/
[2] https://clarin.bbaw.de/bbaw/static/pub/CLARIN_Template_Depositors_Agreement_BBAW.pdf
[3] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page
[4] https://journals.openedition.org/jtei/739
[5] https://clarin.bbaw.de/bbaw/static/pub/Ursel.pdf
[6] https://clarin.bbaw.de/bbaw/static/pub/objectXML.pdf
[7] https://wiki.lyrasis.org/display/FEDORACREATE/Tutorial+1+-+Introduction+to+Fedora#Tutorial1IntroductiontoFedora-Preservation&Archiving

Links:

Reviews

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Set level of compliance to 4 after additional documentation was provided.

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

The repository has a public preservation policy. It is recommended that the described preservation plan is made public before re-certification.

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

ok

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 BBAW CLARIN center repository is integrated into the Common Language Resources and Technology Infrastructure (CLARIN), which implements several channels through which members of the designated communities can give feedback on data and metadata hosted by its certified centres.

The metadata portal CLARIN Virtual Language Observatory [1] harvests the CMDI metadata of all CLARIN centres and displays the large amount of available resources through faceted browsing and search facilities. Both in the overview, i.e. when browsing or searching for relevant resources, and on the individual resource pages displaying further information on a specific resource, the user can report an issue or give feedback on metadata records or resources using a designated button connected via a form to the CLARIN-D Help Desk ticketing system.

The CLARIN-D Help Desk, maintained by the CLARIN centre at the University of Hamburg, 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.

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 [2] that continuously monitors provided metadata of all associated repositories 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).
Further information about the curation module can be found at [3].

At the BBAW CLARIN service center, automated quality checks (e.g. XML validity) are done during the data production/acquisition workflow, see [4].

Part of the archiving workflow is the integrity check of the data and the metadata by the archive manager. This is done both manually and automatically. The metadata is parsed for syntactic correctness and manually evaluated for completeness and soundness. The object data is tested for syntactic correctness if possible.

There is documentation available for

* DTA Basisformat text encoding and metadata, see [5] and [6]

* implemented quality control, see [7]

* the DTAQ collaborative web curator tool, see [8]

* CMDI metadata, see [9].

[1] https://vlo.clarin.eu/
[2] https://curation.clarin.eu/collection/table
[3] https://office.clarin.eu/v/CE-2016-0742-CLARINPLUS-D2_1.pdf
[4] https://clarin.bbaw.de/bbaw/static/img/Archiv_Workflow_en.jpeg
[5] https://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf
[6] https://www.deutschestextarchiv.de/doku/basisformat/introduction_en.html
[7] https://jtei.revues.org/739
[8] https://www.deutschestextarchiv.de/misc/2013-04_poster_allea/poster.pdf
[9] https://www.clarin.eu/cmdi

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:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

ok

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:

With the use of the Fedora-Commons system and the defined workflow supported by the repository's interface, the repository aims to be as conformant to OAIS as possible. Due to the complexity of the OAIS reference model, the repository cannot guarantee that all details are (or will be) implemented. E.g. AIP/DIP/SIPs are not packaged as zip files - the packages are virtual (or better: the files are linked to each other) in our case. The OAIS model basically consists of six functional entities, which we will describe here for the BBAW CLARIN Center:

1. Ingest. This entity receives data from producers. Special tasks are: receiving data, performing quality assurance,checks on documentation, description and formats. Establish metadata and prepare for archiving and data management.

Implications for BBAW CLARIN Center: There is a Standard Operating Procedure for ingest of data (acquisition) which includes all the tasks mentioned.

2. Archival Storage. This entity is responsible for the systematic storage, maintenance and retrieval of the data. It further performs routine checks on media quality (refresh if necessary), errors and disaster recovery capabilities.

Implications for the BBAW CLARIN Center: Two separate functions were implemented: Data management (which is responsible for storage of the data, error detection and retrieval) and system management (which is responsible for media quality and recoverability).

3. Data Management. This entity is responsible for content integrity of the data, version management and theconnection of data and metadata.

Implications for the BBAW CLARIN Center: Content integrity is regularly checked via MD5 checksums. Hard disk deterioration is checked via S.M.A.R.T. status. Version management is achieved via a strict versioning policy, each version is given a handle PID, older versions can be accessed via the object version history. Every dataset needs to have metadata attached, otherwise no ingest is possible due to a workflow restriction.

4. Preservation Planning. This entity is responsible for evaluation of quality of service, state of development intechnology and provides migration planning.

Implications for the BBAW CLARIN Center: The BBAW participates in digital infrastructure projects like CLARIN and Text+ to monitor the technical developments and also community feedback to provide reliable and useful services. The BBAW service center is continuously updated according to the needs of these projects.

5. Administration. This entity is responsible for legal issues like contract agreements and IPR.

Implications for the BBAW CLARIN Center: Before ingest, all data and metadata undergoes a plausibility check to find out whether a valid CC license is attached to the data or if a contract is necessary.

6. Access. This entity is responsible for the interaction with data consumers.

Implications for the BBAW CLARIN Center: Currently all data and metadata are freely available via several interfaces: web frontend with advanced metadata search at [1] , Virtual Language Observatory (VLO) at [2] , CLARIN Federated Content Search (FCS) at [3] and the OAI/PMH-Gateway at [4] .

CLARIN-D has contributed a user-guide [5] which serves as a comprehensive overview on the CLARIN-D infrastructure and describes many best practices used at the service centers.

For the data production/acquisition at the BBAW CLARIN service center, there is documentation available for

* DTA Basisformat text encoding and metadata, see [6] and [7].

* implemented quality control, see [8].

* the DTAQ collaborative web curator tool, see [9]

* CMDI metadata, see [10]

The ingest, management and storage procedures [11] are described in this workflow chart [12].

The online archive management tool Fedora Commons defines a workflow to a certain extent, because no resources can be archived without metadata being present. The depositor determines who can access the material and is also responsible for protecting the privacy of any subjects appearing in the recordings or texts. Additionally quality checks of data and metadata including PID (Persistent Identifier) assignment are done by the repository software.

According to the contract, the depositor may choose between three access levels: 'unrestricted public', 'academic access only' or 'restricted access/only after receiving permission from the depositor'.

Disclosure risk is minimized by anonymization or limited access via login accounts.

In case during the archiving workflow ([13] at the 'DTA inspection and feedback' stage) our staff would find data with disclosure risk, the data would either be rejected until anonymized by the depositor or it would be saved with limited access via login accounts according to the contract. Our staff would provide guidance how anonymization would be done properly.

According to the contract [14] 'the Depositor guarantees that Content contains no data or other elements that are contrary to the law or public regulations'.

Most of the data managed by the BBAW CLARIN repository is text in XML format. These rather small files can be ingested into the Fedora Commons database as 'inline' or 'managed', i.e. Fedora will generate checksums for it.

For larger files (e.g. multimedia) some extra effort is necessary as they have to be stored externally [15].

Change management involves strict versioning and new HANDLE PIDs for the modified data and/or metadata.

[1] https://clarin.bbaw.de/en/search/adv_search/
[2] https://catalog.clarin.eu/vlo
[3] https://www.clarin.eu/contentsearch
[4] https://clarin.bbaw.de:8088/oaiprovider?verb=Identify
[5] https://www.clarin-d.net/en/language-resources-and-services/user-guide
[6] https://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf
[7] https://www.deutschestextarchiv.de/doku/basisformat/introduction_en.html
[8] https://jtei.revues.org/739
[9] https://www.deutschestextarchiv.de/misc/2013-04_poster_allea/poster.pdf
[10] https://www.clarin.eu/cmdi
[11] https://clarin.bbaw.de/en/preservation
[12] https://clarin.bbaw.de/bbaw/static/img/Archiv_Workflow_en.jpeg
[13] https://clarin.bbaw.de/en/repo/#workflow
[14] https://clarin.bbaw.de/bbaw/static/pub/CLARIN_Template_Depositors_Agreement_BBAW.pdf
[15] https://wiki.duraspace.org/display/FEDORA38/Fedora+Digital+Object+Model#FedoraDigitalObjectModel-Datastreamsdata

Links:

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:

The guideline has been fully implemented in the repository - 4

Comments:

ok

R13 Data discovery and identification

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The repository provides various ways of utilizing the archived data via online tools as well as by downloading the data in formats commonly used by the research communities. An advanced metadata search utility [1] is provided, as well as a simple search tool [2].

Additionally, CLARIN provides search facilities like the Virtual Language Observatory VLO [3] to lookup digital assets in all CLARIN center repositories and the Federated Content Search Aggregator FCS [4] to enable full-text search across all text corpora available in all CLARIN centers.

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

The repository enables metadata harvesting via OAI/PMH protocol and the metadata formats Dublin Core (DC) and Component MetaData Infrastructure (CMDI [5]).

OAI-PMH endpoint [6]

The repository is listed in the following registries:

* OpenDOAR [7]

* Re3Data [8]

* CLARIN Centre Registry [9]

* Duraspace Registry [10]

The repository offers recommended data citations at the bottom of each search record page (including the CMDI metadata PID and attribution to individuals/organizations who contributed to their creation).

The repository itself does not offer a persistent identifier service on it's own but makes use of a common CLARIN PID service [11] based on the handle system [12], in cooperation with the European Persistent Identifier Consortium (EPIC [13]). The usage of PIDs is mandatory for resources in CLARIN, thus all resources added to the repository may be referenced using PIDs. The PIDs are defined according to ISO 24619:2011.

[1] https://clarin.bbaw.de/en/search/adv_search
[2] https://clarin.bbaw.de/en/search/
[3] https://vlo.clarin.eu
[4] https://www.clarin.eu/contentsearch/
[5] https://www.clarin.eu/cmdi
[6] https://clarin.bbaw.de:8088/oaiprovider?verb=Identify
[7] https://v2.sherpa.ac.uk/id/repository/3986
[8] https://www.re3data.org/repository/r3d100012054
[9] https://centres.clarin.eu/centre/6
[10] https://duraspace.org/registry/entry/3308/
[11] https://www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf
[12] https://www.handle.net/
[13] https://www.pidconsortium.eu/

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

ok

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:

For metadata we especially rely on the CMDI [1] format which is more expressive and flexible than Dublin Core(DC [2]). CMDI metadata sets are made human readable by the use of XSL stylesheets in the repository.

All CLARIN centres [3] provide their metadata in the CMDI format. The Component MetaData Infrastructure (CMDI [1]) was initiated by CLARIN to provide a flexible framework for describing metadata based on components and concepts. Each metadata record is based on a profile that is registered in the CLARIN CMDI Component Registry [4]. Profiles can make use of components. Those building blocks are also registered in the CMDI Component Registry and describe specific aspects or properties of a resource. Elements of CMDI records link to concept definitions that are stored in external registries (like the CLARIN Concept Registry [5]). Since different communities use different names for the same concepts, linking CMDI elements to concepts enables communities to stick to their terminology while enabling users to find concepts independent of the naming.

Another strict requirement for CLARIN centres is to make their metadata also available through the established and well documented Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH [6]). This standard enables harvesting of the metadata from the repository via HTTP(S).

In 2017 our repository made a transition from CMDI metadata version 1.0 to 1.2. This has been achieved by a standard XSL transformation of previous CMDI metadata versions. Future migrations are expected to be handled in a similar way.

Measures are taken to ensure the future interpretability of the data. The number of accepted file formats is limited, to make future conversions to other formats more feasible. Open (non-proprietary) file formats are used whenever possible. For textual resources, XML formats are used whenever possible, to ensure future interpretability of the files independent of the tool used to create them. Text is encoded in Unicode to ensure future interpretability.

Understandability of the data and metadata is covered in multiple ways. For data, we use the well-documented XML standard of the Text Encoding Initiative (TEI [7]).

The structural annotation of all texts is done according to the DTA 'base format' (DTABf [8]). The DTABf was developed as a subset of the P5-Guidelines [9] of the Text Encoding Initiative (TEI [7]) by the team of the Deutsches Textarchiv (DTA [10]). It is widely used in the community and recommended as a standard format by the Deutsche Forschungsgemeinschaft (DFG [11]) and the CLARIN-D User Guide [12] . Since the TEI Guidelines are offering solutions for a huge amount of tagging requirements and are thus rather extensive and flexible, they are meant to be adjusted to the individual necessities of projects working with the TEI. For the DTA this was achieved by creation of the DTABf, a proper subset of the TEI/P5 tag set [9] which offers not only fixed sets of elements but also of corresponding attributes and (where applicable) values. The DTABf tagset is fully conformant with the TEI/P5-Guidelines, i.e. the TEI tag set was only reduced not extended in any way.

A document about the scope of data curation at the BBAW CLARIN center repository as well as collection development and the minimum set of metadata is available here [13].

All points of the bullet list in the document linked above are collected when curating new data.

We're using the CMDI profile clarin.eu:cr1:p_1381926654438. It consists of the following components:

Component Name Description ID

* author the author of a textual resource clarin.eu:cr1:c_1493735943963
* editionStmt describes the edition of a resource clarin.eu:cr1:c_1342181139673
* editor the editor of a textual resource clarin.eu:cr1:c_1396012485117
* extent the size of a resource with respect to a specified unit of measurement clarin.eu:cr1:c_1345180279117
* fileDesc metadata for the electronic edition of a text clarin.eu:cr1:c_1381926654441
* idno known IDs for a specific entity clarin.eu:cr1:c_1493735943964
* orgName an organization's name clarin.eu:cr1:c_1381926654512
* persName a person's name clarin.eu:cr1:c_1493735943962
* profileDesc detailed description of non-bibliographic aspects of a text clarin.eu:cr1:c_1493735943961
* publicationStmt administrative information regarding the publication of a resource clarin.eu:cr1:c_1381926654439
* seriesStmt information about book series and journals clarin.eu:cr1:c_1498745062851
* sourceDesc bibliographical and physical properties of a source text clarin.eu:cr1:c_1381926654443
* titleStmt bibliographical information on author or editor and title of a text clarin.eu:cr1:c_1381926654513

The maximum of metadata currently used (more comprehensive, possibly expected documentation by the repository) can be seen here (in the chapter: "Annotation of Metadata") [14].

English translation of DTA Basisformat documentation: [15]

[1] https://www.clarin.eu/cmdi
[2] https://dublincore.org/
[3] https://www.clarin.eu/content/overview-clarin-centres
[4] https://catalog.clarin.eu/ds/ComponentRegistry
[5] https://openskos.meertens.knaw.nl/ccr/browser/
[6] https://www.openarchives.org/pmh/
[7] https://www.tei-c.org/
[8] https://hdl.handle.net/21.11120/0000-0000-F484-7
[9] https://www.tei-c.org/Guidelines/P5/
[10] https://www.deutschestextarchiv.de/
[11] https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literat
[12] https://media.dwds.de/clarin/userguide/text/text_corpora.xhtml
[13] https://clarin.bbaw.de/en/curation/#DTAE_Checklist
[14] https://hdl.handle.net/21.11120/0000-0000-F4B5-0
[15] https://www.deutschestextarchiv.de/doku/basisformat/introduction_en.html

Links:

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:

The guideline has been fully implemented in the repository - 4

Comments:

ok

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 Operating System used by the BBAW CLARIN repository is Debian GNU/Linux ([1] 'stable' release with 5 years long term support (LTS), which enables to upgrade to the next releases easily). In accordance with the BBAW internal IT security concept (in german), all operating systems are patched at least once per month. Distribution upgrades to the latest stable release are performed before support is running out.

The repository web frontend for Fedora Commons was developed based on the web framework Django [2], EULFedora libraries [3] and a MySQL database. Our 'Fedora handler client' software is an in-house development which is written in Java. It performs operations such as ingest and updating Fedora records by calling Fedora REST-API functions [4]. Our OAI gateway software is community based [5].

Backups are performed when the data in the repository changes, and are stored in the form of disaster recoverable virtual machine images as well as file system and database dumps. Virtual machine backups can be immediately restarted on other virtualization hardware which is in place in the secondary server room.

For software backups, we dump databases to local storage, sync those dumps (via rsync [6]), and additionally sync local software daily to a another server. Weekly backups are performed to a tape library via the backup software Bareos [7], which determines independently when incremental and full dumps have to be made (but full dumps are done at least once per month). Bareos is an open source software fork of the popular software Bacula [8].

In addition to software backups, the virtual machines are completely backed up as virtual machine image snapshots via Proxmox vzdump [9], which are themselves then backed up to tape storage to ensure fast disaster recovery times and facilitate live migration of virtual machines to another virtualization cluster node. Proxmox uses the open source linux kernel virtual machine (kvm) software internally, which again ensures the ability to recover or convert snapshots also in the distant future. The snapshots are performed prior to configuration updates on the machines. In worst case the content of the repository could be transferred to another CLARIN center, especially making use of persistent identifiers (PIDs) and backup copies.

With the use of the Fedora-Commons system [10] and the defined workflow supported by the repository's interface, the repository aims to be as conformant to OAIS as possible.

CLARIN-D has contributed a user-guide [11] which serves as a comprehensive overview on the CLARIN-D infrastructure and describes many best practices used at the service centers.

For metadata we rely on the group of emerging standards around CMDI (ISO-CD 24622-1) [12] .

For data, we use the well-documented XML standard of the Text Encoding Initiative (TEI). The structural annotation of all texts is done according to the DTA 'base format' (DTABf [13]). The DTABf was developed in accordance with the P5-Guidelines of the Text Encoding Initiative (TEI [14]). It is widely used in the community [15] and recommended as a standard format by the Deutsche Forschungsgemeinschaft (DFG [16]) and the CLARIN-D User Guide . Since the TEI Guidelines are offering solutions for a huge amount of tagging requirements and are thus rather extensive and flexible, they are meant to be adjusted to the individual necessities of projects working with the TEI. For the DTA this was achieved by creation of the DTABf, a proper subset of the TEI/P5 tag set [17] , which offers not only fixed sets of elements but also of corresponding attributes and (where applicable) values. The DTABf tag set is fully conformant with the TEI/P5-Guidelines, i.e. the TEI tag set was only reduced not extended in any way.

DTA Basisformat text encoding and metadata, see [18] and [19].

Important goals of infrastructure development are [20] :

* 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.

There is software documentation internally available for cases of emergency (what to check if the server isn't working properly, server and software dependencies, installation guidelines and staff to contact). A software inventory is available internally (generated via the Debian Linux package repository).

The BBAW CLARIN repository is hosted on two virtualization servers in two different server rooms (the main data center and a backup server room) at the BBAW. The server rooms are in different fire safety zones. Both server rooms have redundant cooling and redundant uninterruptible power supplies. A duplicate of the virtual machine backup can be started in the secondary server room in case of a disaster for swift recovery within minutes.

Access to the data center is limited to authorized staff. Maintenance of the systems is performed by a professional systems administrator.

Access to the virtual server is restricted by a firewall. The storage hardware and hardware for virtual machines is replaced at regular intervals to the latest state of art.

The BBAW CLARIN repository virtual machine, the backup server and other critical infrastructure is monitored with Icinga (= network and service monitoring software).

The repository hasn't ingested real-time stream data yet, the network connectivity of the BBAW building is provided by two redundant connections (each 1GBit/s bandwidth download, 1GBit/s bandwidth upload) by two different carriers on different routes though. Both connections are bundled via Link Aggregation (so we use 2 GBit/s download, 2 GBit/s upload effectively). For the hardware we rely on two virtualization servers (based on Proxmox Linux [21]) and storage systems which are configured for failover and connected to each other via 10GBE fiber network in different fire safety zones (in german 'Brandschutzabschnitt') of the building.

[1] https://www.debian.org/
[2] https://www.djangoproject.com/
[3] https://eulfedora.readthedocs.org
[4] https://wiki.duraspace.org/display/FEDORA38/REST+API
[5] http://proai.sourceforge.net/
[6] https://rsync.samba.org/
[7] https://www.bareos.org
[8] https://www.bacula.org
[9] https://pve.proxmox.com/wiki/Backup_and_Restore
[10] https://fedorarepository.org/
[11] https://www.clarin-d.net/en/language-resources-and-services/user-guide
[12] https://www.clarin.eu/cmdi
[13] https://www.deutschestextarchiv.de/doku/basisformat/introduction_en.html
[14] https://www.tei-c.org/
[15] https://www.deutschestextarchiv.de/clarin-kooperationen
[16]
https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literat
[17] https://www.tei-c.org/Guidelines/P5/
[18] https://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf
[19] https://www.deutschestextarchiv.de/doku/basisformat/introduction_en.html
[20] https://www.clarin.eu/content/clarin-technology-introduction
[21] https://www.proxmox.com/en/

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:

ok

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:

Entry to the building is restricted by security staff. BBAW visitors use a separate wifi network (EDUROAM), so there is a seggregation betweeen visitors and BBAW staff.

Servers and services are monitored by a local Icinga monitoring system, but also externally from the CLARIN monitoring system [1].

There is software documentation internally available for cases of emergency (what to check if the server isn't working properly, server and software dependencies and staff to contact).

There are three levels of security for data access required - public open access (no authentication required), research community open access (authentication via Shibboleth is required) and restricted access (authentication via personal login account).

An authentication and authorization infrastructure (the CLARIN service provider federation, aiming to include all european identity provider federations, based on Shibboleth) is available and in use in the repository for use of "academic only" resources [2] .

The BBAW appointed an IT security officer and his representative in 2013. In 2014 an internal IT security concept (in german) was developed in consultation with the Head of IT, the data security officer and the legal advisor of the BBAW. It aims at being compliant with the comprehensive BSI-IT-Grundschutz [3] developed by the Federal Office for Information Security in Germany (BSI [4]). The IT security concept has been finalized and covers topics like secure networking, backups and archiving, antivirus protection, encryption, patch management policy, user management, desktop and server security and so on.

In preparation of a new data center build for the BBAW in 2010, a consulting company for high availability server rooms and risk management evaluated the current situation and gave us advice for the setup of the data center especially concerning fire safety, redundant cooling and redundant uninterruptible power supplies. Also they recommended us to have a fallback server room in another fire safety zone. This has been implemented, so today we have virtualization hardware present in two server rooms in different fire safety zones so we can start a duplicate of the virtual machine backup in the secondary server room in case of a disaster within minutes. Also the network connection to the building is redundant via two independent carriers. Air conditioning and uninterruptible power supplies have also been implemented redundantly (air conditioning is powered by uninterruptable power supplies, too). The main data center has early fire detection and fire suppression system using argon gas as suppression agent.

In 2021 the BBAW exchanged their firewall infrastructure to a redundant, dual 'next generation firewall' (NGFW) setup including intrusion prevention & detection mechanisms with a demilitarized zone (DMZ) for the servers.

[1] https://status.clarin.eu/
[2] https://www.clarin.eu/content/service-provider-federation
[3] https://www.bsi.bund.de/EN/Topics/ITGrundschutz/itgrundschutz_node.html
[4] https://www.bsi.bund.de/EN/

Links:

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:

ok

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 your valuable feedback!

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:

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments: