Implementation of the CoreTrustSeal

The CoreTrustSeal board hereby confirms that the Trusted Digital repository CLARIN Center BBAW complies with the guidelines version 2017-2019 set by the CoreTrustSeal Board.
The afore-mentioned repository has therefore acquired the CoreTrustSeal of 2016 on October 25, 2018.

The Trusted Digital repository is allowed to place an image of the CoreTrustSeal logo corresponding to the guidelines version date on their website. This image must link to this file which is hosted on the CoreTrustSeal website.

Yours sincerely,

The CoreTrustSeal Board

Assessment Information

Guidelines Version:	2017-2019 | November 10, 2016
Guidelines Information Booklet:	CTS Requirements 2017-2019
All Guidelines Documentation:	Documentation
Repository:	CLARIN Center BBAW
Seal Acquiry Date:	Oct. 25, 2018
For the latest version of the awarded CoreTrustSeal for this repository please visit:	http://www.coretrustseal.org/why-certification/certified-repositories/
Previously Acquired Seals:	Seal date: May 8, 2015 Guidelines version: 2014-2017 | July 19, 2013 Seal date: May 21, 2013 Guidelines version: 2010 | June 1, 2010
This repository is owned by:	**Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)** Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) Jägerstr. 22-23 Zentrum Sprache 10117 Berlin Germany Berlin Germany T +49 (0)30 20370 0 F +49 (0)30 20370 600 E clarin@bbaw.de W http://www.bbaw.de/

Assessment

0. Context

Applicant Entry

Self-assessment statement:

The Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) 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 in its CLARIN-D service center at the BBAW publishes and preserves historical and contemporary German text corpora as well as the lexical resources provided by the Zentrum Sprache (Language Centre) at the BBAW - mostly in open formats like XML. It is part of CLARIN-D (Common Language Resources and Technology Infrastructure Deutschland) - a web and centers-based research infrastructure for the social sciences and humanities. The aim of CLARIN-D and its service centers is to provide language data, tools and services in an integrated, interoperable and scalable infrastructure for the social sciences and humanities. The research infrastructure is rolled out in close collaboration with expert scholars in the humanities and social sciences, to ensure that it meets the needs of users in a systematic and easily accessible way. CLARIN-D is funded by the German Federal Ministry for Education and Research.

CLARIN-D collaborates with and uses infrastructural components of the European CLARIN (https://www.clarin.eu/) initiative. Research standards to be met by the CLARIN services centers, technical standards and solutions for key functions, a set of requirements which participating centers must provide, as well as plans for the sustainable provision of tools and data and their long-term archiving have been developed.

According to CLARIN terminology this is a resource center certified as type B. CLARIN distinguishes a number of different center types that have different impact for the language resources and tools infrastructure. Type B centers 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.

Within CLARIN-D the following requirements hold for centers of type B (https://www.clarin.eu/node/3542) and are fulfilled by this resource center:

Centers need to offer useful services to the CLARIN community and to agree with the basic CLARIN principles (own architecture choice, explicit statement about quality of service, usage of persistent identifiers, adherence to agreed formats, protocols and APIs).

Centers need to adhere to the security guidelines, i.e. the servers need to have accepted certificates.

Centers 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 Security Assertion Markup Language (SAML2.0) and trust declarations. In case all resources at a center are open, setting up a Service Provider is optional.

Centers need to have a proper and clearly specified repository system and participate in a quality assessment procedure as proposed by the Data Seal of Approval or MOIMS-RAC approaches.

Centers need to offer component based metadata (CMDI) that make use of elements from accepted registries such as CLARIN Concept Registry in accordance with the CLARIN agreements, i.e. metadata needs to be harvestable via OAI PMH.

Centers need to associate PID records according to the CLARIN agreements with their objects and add them to the

metadata record.

Each center needs to make clear statements about their policy of offering data and services and their treatment of IPR (intellectual property rights) issues.

Each center needs to make explicit statements to the CLARIN boards about its technological and funding support state and its perspectives in these respects.

Centers need to employ activities to relate their role in CLARIN to the research community in order to guarantee a research based status of the infrastructure and allow researchers to embed their services in their daily research work.

Centers that are offering infrastructure type of services need to specify their services for CLARIN and the terms of giving service.

Centers are advised to participate in the Federated Content Search with their collections by providing an SRU/CQL Endpoint. This content search is especially suitable for textual transcriptions and resources.

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

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), these datasets are curated on data-level (level D) by their team at the BBAW CLARIN center.

List of outsource partners:

1) Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
The repository makes use of a common CLARIN PID service (https://www.clarin.eu/sites/default/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. The services provided are stipulated as follows:

GWDG provides PID services for eight prefixes for up to 50.000 PIDs per year and two PID services for up to a million PIDs per year. These PID services based on the handle system include:

A service for minting PlDs via EPIC API v2 with its current releases during the period of contract.

All minted PIDs are additionally twofold replicated to mirror servers of the EPIC consortium.

A permanent resolution service of minted PIDs is goal of the EPIC consortium. A resolution service for at least ten years is guaranteed by GWDG.

The offer distinguishes base costs for these services, that include the minting of up to 50.000 PIDs per year and additional costs, that include the minting of up to 1.000.000 PIDs per year.

This does not include software development for special requirements of the project.

This outsource partner offers relevant functionality for guideline 13: „The data repository enables the users to discover the data and refer to them in a persistent way through proper citation.“.

Being a CLARIN-D centre, we're making use of CLARIN infrastructure services (which are considered as outsourced). This applies to the following relevant services:

- Authentication and Authorization Infrastructure 'AAI identity federation' for Single Sign On (https://www.clarin.eu/content/federated-identity driven by CLARIN / CLARIN-ERIC)
- Virtual Language Observatory 'VLO' for harvesting and searching metadata (https://vlo.clarin.eu driven by CLARIN / CLARIN-ERIC)
- Federated Content Search 'FCS' (https://www.clarin.eu/content/content-search driven by CLARIN / CLARIN-ERIC)
- multiple Registries for managing metadata (driven by CLARIN / CLARIN-ERIC)
- Helpdesk for managing support and feedback workflows (https://clarin.bbaw.de/en/kontakt/ driven by CLARIN-D)
- transfer of repository content to another center in case a center ceases to exist (driven by CLARIN-D)

see also https://www.clarin.eu/value-proposition (especially the linked PDF document http://hdl.handle.net/11372/DOC-138 on that page)

and https://www.re3data.org/repository/r3d100012054

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

1. Mission/Scope

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

The mission of this repository is to ensure the availability and long-term 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). 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 and has been officially enacted by the representative (president) of the BBAW.

see https://clarin.bbaw.de/en/mission

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

2. Licenses

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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, license types, compensation/payments, liability for damages and costs, termination of the agreements

An example contract can be downloaded here:

http://clarin.bbaw.de/bbaw/static/pub/CLARIN_Template_Depositors_Agreement_BBAW.pdf

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

Most of the data in the repository have Creative Commons licenses (https://creativecommons.org) applied to them (unrestricted access). 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):

https://www.clarin.eu/content/federated-identity

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

We specifically rely on 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.
http://www.allea.org/Content/ALLEA/Scientific%20Integrity/Code_Conduct_ResearchIntegrity.pdf

DFG, Rules of Good Scientific Practice
http://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice/

BBAW, Richtlinien zur Sicherung guter wissenschaftlicher Praxis
http://www.bbaw.de/die-akademie/aufgaben-und-ziele/sicherung-guter-wissenschaftlicher-praxis/RichtlinienundAusfuehrungsbestimmun

Data users have to follow the repository's General Terms of Use. It is linked in the data catalog at the bottom of each bibliography page and also in the repository description page.

see also:

https://clarin.bbaw.de/en/repo/

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

3. Continuity of access

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

3. In progress: We are in the implementation phase.

Self-assessment statement:

As stated in the Depositor's Agreement, 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.

The CLARIN-D centres have signed a memorandum of understanding 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.

For the BBAW CLARIN centre and its repository, funding is secured until 09/2024.

The Center staff consist of part-time appointees such that the work force sums up to 4 FTE positions: 2 FTE are financed by CLARIN-D (funded by BMBF, secured until 09/2020) and 2 FTE are financed by BBAW (funded by the umbrella organization of the German academies, secured until 09/2024). In case the CLARIN-D funding wouldn't continue, BBAW would continue maintaining the repository at least until 09/2024.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

4. Confidentiality/Ethics

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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'.

http://clarin.bbaw.de/bbaw/static/pub/CLARIN_Template_Depositors_Agreement_BBAW.pdf

In case during the archiving workflow (https://clarin.bbaw.de/en/repo/#workflow 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.

Ethical rules

ALLEA (ALL European Academies) European Science Foundation, The European Code of Conduct for Research Integrity. http://www.allea.org/Content/ALLEA/Scientific%20Integrity/Code_Conduct_ResearchIntegrity.pdf

DFG, Rules of Good Scientific Practice,
http://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice/

BBAW, Richtlinien zur Sicherung guter wissenschaftlicher Praxis,
http://www.bbaw.de/die-akademie/aufgaben-und-ziele/sicherung-guter-wissenschaftlicher-praxis/RichtlinienundAusfuehrungsbestimmun

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

5. Organizational infrastructure

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

The Berlin-Brandenburg academy of sciences and humanities (founded in 1700) is a recognized institution for longterm projects. For the BBAW CLARIN service center, its repository and staff, funding is secured until 09/2024.

In addition our repository is part of CLARIN-D, a research infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences. CLARIN-D also offers information on a wide range of topics, including teaching material, help on data management plans and other, discipline-specific support. By being part of the CLARIN-D consortium the repository gains access to funding for running and further developing a sustainable repository and resource center to support these goals.

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 Center staff consist of part-time appointees such that the work force sums up to 4 FTE positions: 2 FTE are financed by CLARIN-D (funded by BMBF, secured until 09/2020) and 2 FTE are financed by BBAW (funded by the umbrella organization of the German academies, secured until 09/2024). In case the CLARIN-D funding wouldn't continue, BBAW would continue maintaining the repository at least until 09/2024.

Administration
Project leader, Board reporting (Computational Linguist)

Technology
Software Developer, web administration (IT engineer)
Software Developer, linguistic tools and search engines infrastructure (Computational Linguist)
Systems Administrator, Systems, networking, hardware and software, ingest of data (IT specialist)

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

The repository staff members have access to training on data management, metadata, long-term preservation and professional development (offered by CLARIN-D and CLARIN-ERIC). This includes budget for regular developer meetings, mobility grants for sharing of expertise, conferences, workshops, meetings with their

respective scientific communities (called discipline-specific working groups, https://www.clarin-d.net/en/disciplines) as well as a centralized knowledge base (user guide, wiki, bugtracker and mailing lists).

Currently CLARIN-D is funded by the Bundesministerium für Bildung und Forschung (BMBF). The current project phase has a runtime of 4 years and is funded until 30.09.2020. As an alternative to project based funding, CLARIN-D currently pursuits a permanent continuation of funding.

CLARIN has a wide field of expertise in its collaborative network of centers, which come from within their respective fields of digital humanities.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

6. Expert guidance

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

Most of the BBAW repository content is provided by the Deutsches Textarchiv (http://www.deutsches-textarchiv.de). The texts were selected on the basis of a comprehensive bibliography of influential books prepared by members of the BBAW as well as specialists in various disciplines.

CLARIN-D is supported by external advisory committees. The International Advisory Board (IAB), CLARIN-D's scientific advisory board, is a group of CLARIN-D external experts who are consulted on new developments and discuss strategic and content related developments, also with a bird-eye view of other developments in the communities. With experienced experts from various backgrounds, a high-profile international committee was formed for this purpose.

The joint Technical Advisory Board (TAB) of CLARIN-D and DARIAH-DE is a committee supports collaboration on the fundamental technical level between two large research infrastructures for the humanities and social sciences. The issues of the Collaboration are: questions of technical protocols, infrastructural requirements on the level of archiving, interconnection, search, etc. Based on requirements, small working groups (for example on persistent identifiers, authorization and identification) are being formed in areas with an overlap of requirements. This avoids duplication of developments and allows an increased efficiency in implementation, but also interoperability where overlaps exist. This includes for example an option to grant access to one infrastructure for users of the other.

CLARIN is committed to boosting humanities research in a multicultural and multilingual Europe, by facilitating access to language resources and technology for researchers and scholars across a wide spectrum of domains in the humanities and social sciences (HSS). To reach this goal and to contribute to overcome the traditional gap between the Humanities and the Language Technology communities we established an active interaction with the research communities in HSS in so called discipline-specific working groups (https://www.clarin-d.net/en/disciplines).

These groups act as a link between the CLARIN-D resource centres and the research communities which represent the users of the CLARIN-D infrastructure. Currently eight working groups act as consultants for the needs of the humanities, social sciences and particular disciplines. All together they consist of more than 100 academic professionals. Their main role is to advise CLARIN-D during the development and implementation of the infrastructure so that these efforts can best meet the needs of all research communities involved. The working group chairs further coordinate dissemination and best practice using CLARIN-D services in their member communities.

CLARIN-D organizes joint activities of the working groups. This includes the organization of working group meetings, organization of specialized and interdisciplinary workshops and the creation of joint reports. Further, communications between CLARIN-D centres and the working groups as well as groups among themselves are coordinated. Virtual meetings are held on a monthly basis. Contents of the curation projects and activities of the WG are published on the CLARIN-D Website (https://www.clarin-d.net/en/disciplines/). For communication, mailing lists and wiki contents are maintained.

At the BBAW CLARIN center, we're in direct contact with historians, lexicographers, psychologists, literary scholars and germanists (specialists in German studies). For these groups our text corpora and lexical resources are particularly relevant. E.g. in the design phase of reference corpora these groups are asked for relevant text sources.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

7. Data integrity and authenticity

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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 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 and DC format) also is present.

Most of the data in the repository is ingested by the Deutsches Textarchiv (DTA) which uses version control mechanisms (provided by GIT version control software) to log changes. The repository itself currently does not log differences in metadata or data.

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, '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 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.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

8. Appraisal

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

A collection development/appraisal document describing the scope of data curation at the BBAW CLARIN center repository is available here: https://clarin.bbaw.de/en/curation/#Contents

Generally the repository accepts historical and contemporary German text corpora as well as the lexical resources provided by the Zentrum Sprache (Language Centre), but the Deposits are handled in a case-by-case approach.

Quality checks are done before ingest by our staff. See the workflow image.

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

http://hdl.handle.net/21.11120/0000-0000-F4B5-0

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. Lists of recommended formats CLARIN standards recommendations: http://www.clarin.eu/recommendations

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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

9. Documented storage procedures

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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 (http://clarin.bbaw.de/en/documentation/#Storage_Procedures):

"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-5 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.

LTO tape media is periodically tested for deterioration by reading the error information from its memory chip (LTO-CM). Also the backup software internally makes use of checksums to recognize tape block errors.

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."

Generally data access is governed by our Open Access policy which is in line with the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities.

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).

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.

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). Metadata is a data stream within the digital object, and as such is version controlled like object data. CLARIN propagates the idea of reproducible research. Thus updates/new versions of resources typically are equipped with a new PID.

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. All datastreams and versions are equipped with a MD5 checksum, which is checked in coordination with the backups as described above.

For further details on the backup part of the workflow see also R16.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

10. Preservation plan

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

3. In progress: We are in the implementation phase.

Self-assessment statement:

We indicate to Data Producers and to our community (via our mission statement) 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.

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.

The example depositor's contract (http://clarin.bbaw.de/bbaw/static/pub/CLARIN_Template_Depositors_Agreement_BBAW.pdf) provides for all actions to meet the responsibilities (licensing, copying, storing, modification, distribution).

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

Currently we're working on a preservation plan which will be published on the repository website (and will be added in the next review).

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

11. Data quality

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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 (https://vlo.clarin.eu) 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.

Furthermore, the so-called discipline-specific working groups within the CLARIN-D project (https://www.clarin-d.net/en/disciplines) 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.

At the BBAW CLARIN service center, automated quality checks (e.g. XML validity) are done during the data production/acquisition workflow, see https://clarin.bbaw.de/bbaw/static/img/Archiv_Workflow_en.jpeg

There is documentation available for

DTA Basisformat text encoding and metadata, see
http://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf

http://www.deutschestextarchiv.de/doku/basisformat/introduction_en.html

implemented quality control, see:
http://jtei.revues.org/739

the DTAQ collaborative web curator tool, see:
http://www.deutschestextarchiv.de/misc/2013-04_poster_allea/poster.pdf

CMDI metadata, see:
https://www.clarin.eu/cmdi

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

12. Workflows

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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 the connection 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 in technology and provides migration planning.

Implications for the BBAW CLARIN Center: The BBAW participates in digital infrastructure projects like CLARIN and DARIAH 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 https://clarin.bbaw.de/en/search/adv_search/ , Virtual Language Observatory (VLO) at http://catalog.clarin.eu/vlo , CLARIN Federated Content Search (FCS) at https://www.clarin.eu/contentsearch and the OAI/PMH-Gateway at https://clarin.bbaw.de:8088/oaiprovider?verb=Identify.

CLARIN-D has contributed a user-guide (https://www.clarin-d.net/en/language-resources-and-services/user-guide) 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:
http://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf
http://www.deutschestextarchiv.de/doku/basisformat/introduction_en.html

implemented quality control, see: http://jtei.revues.org/739

the DTAQ collaborative web curator tool, see:
http://www.deutschestextarchiv.de/misc/2013-04_poster_allea/poster.pdf

CMDI metadata, see: https://www.clarin.eu/cmdi

The ingest, management and storage procedures are described in this workflow chart:
http://clarin.bbaw.de/en/documentation
https://clarin.bbaw.de/bbaw/static/img/Archiv_Workflow_en.jpeg

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 for the depositor'.

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

In case during the archiving workflow (https://clarin.bbaw.de/en/repo/#workflow 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 'the Depositor guarantees that Content contains no data or other elements that are contrary to the law or public regulations'.

http://clarin.bbaw.de/bbaw/static/pub/CLARIN_Template_Depositors_Agreement_BBAW.pdf

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. https://wiki.duraspace.org/display/FEDORA38/Fedora+Digital+Object+Model#FedoraDigitalObjectModel-Datastreamsdata

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

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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

13. Data discovery and identification

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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 (http://clarin.bbaw.de/en/search/adv_search) is provided, as well as a simple search tool (http://clarin.bbaw.de/en/search/).

Additionally, CLARIN provides search facilities like the Virtual Language Observatory VLO (http://www.clarin.eu/vlo/) to lookup digital assets in all CLARIN center repositories and the Federated Content Search Aggregator FCS (https://www.clarin.eu/contentsearch/) to enable full-text search across all text corpora available in all CLARIN centers.

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

OAI-PMH endpoint:

https://clarin.bbaw.de:8088/oaiprovider?verb=Identify

The repository is listed in the following registries:

http://v2.sherpa.ac.uk/id/repository/3986 (OpenDOAR)

https://www.re3data.org/repository/r3d100012054

https://centres.clarin.eu/centre/6

https://duraspace.org/registry/entry/3308/

The repository offers recommended data citations at the bottom of each search record page (including the CMDI metadata PID).

The repository itself does not offer a persistent identifier service on it's own but 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/), 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. The PIDs are defined according to ISO 24619:2011.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

14. Data reuse

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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

All CLARIN centres (https://www.clarin.eu/content/overview-clarin-centres) provide their metadata in the CMDI format. The Component MetaData Infrastructure (CMDI) 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.

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). This standard enables harvesting of the metadata from the repository via HTTP(S).

Currently our repository makes a transition from CMDI metadata version 1.0 to 1.2. This has been done for the majority of our records and is 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). The structural annotation of all texts is done according to the DTA 'base format' (DTABf). The DTABf was developed as a subset of the P5-Guidelines of the Text Encoding Initiative (TEI) by the team of the Deutsches Textarchiv (DTA). It is widely used in the community and recommended as a standard format by the Deutsche Forschungsgemeinschaft (DFG) 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, 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:
https://clarin.bbaw.de/en/curation/#DTAE_Checklist
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"):
http://hdl.handle.net/21.11120/0000-0000-F4B5-0

english translation of DTA Basisformat documentation:
http://hdl.handle.net/21.11120/0000-0000-F484-7

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

15. Technical infrastructure

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

The Operating System used by the BBAW CLARIN repository is Debian GNU/Linux (see: https://www.debian.org/ ; '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 (https://www.djangoproject.com/), EULFedora libraries (http://eulfedora.readthedocs.org) 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 (see https://wiki.duraspace.org/display/FEDORA36/REST+API). Our OAI gateway software is community based (see http://proai.sourceforge.net/).

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, http://rsync.samba.org/), and additionally sync local software daily to a another server. Weekly backups are performed to a tape library via the backup software Bareos (see http://www.bareos.org), 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 (http://www.bacula.org).

In addition to software backups, the virtual machines are completely backed up as virtual machine image snapshots via Proxmox vzdump (see https://pve.proxmox.com/wiki/Backup_and_Restore), 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.

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.

CLARIN-D has contributed a user-guide (https://www.clarin-d.net/en/language-resources-and-services/user-guide) 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), see http://www.clarin.eu/cmdi

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). The DTABf was developed in accordance with which is based on the P5-Guidelines of the Text Encoding Initiative (TEI). 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, 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:

http://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf
http://www.deutschestextarchiv.de/doku/basisformat/introduction_en.html

As part of CLARIN-D we are committed to play an active role in the development of CLARIN's repository infrastructure. General plans for maintaining and further developing the infrastructure have been formulated as part of the project proposal.
The central goal is to improve the usability of the research infrastructure for typical research tasks such as the retrieval of resources, the evaluation of data or the publication of results. To achieve this, modifications and extensions to a variety of infrastructure components in the repository and in the central infrastructure are necessary. Meetings of all centres to monitor advances in infrastructure development take place quarterly.
Further important goals of infrastructure development are
(https://www.clarin.eu/content/clarin-technology-introduction):

- To ensure resilience, integrity, and availability of the sustainable repositories and the central infrastructure

- To integrate new resources and tools based on the needs of the user communities

- To allow for better interoperability of tools and resources in the infrastructure

- To enhance the central content search to be more useful in actual research tasks

- To optimize metadata of the resources provided and to enhance user experience in central metadata search

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

There is documentation available from CLARIN-D on how to setup Fedora Commons (http://www.clarin-d.de/images/leipzig/Fedora_OAI_Konfiguration_v3.pdf). Also 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). 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 redundant cooling and redundant uninterruptible power supplies. The main data center has early fire detection and fire suppression system using argon gas as suppression agent. A duplicate of the virtual machine backup can be started in the secondary server room in case of a disaster within minutes.

Access to the data center is limited to authorized staff. Maintenance of the systems is performed by a professional systems administrator (full-time position).

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 upload, 600MBit/s bandwidth download) by two different carriers on different routes though.

For the hardware we rely on two highly available virtualization servers (based on Proxmox Linux) 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.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

16. Security

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

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).

Servers and services are monitored by a local Icinga monitoring system, but also externally from the CLARIN monitoring system (http://status.clarin.eu/).

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 (https://www.bsi.bund.de/EN/Topics/ITGrundschutz/itgrundschutz_node.html) developed by the Federal Office for Information Security in Germany (BSI). The IT security concept is work in progress 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.

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, http://rsync.samba.org/), and additionally sync local software daily to a another server. Weekly backups are performed to a tape library via the backup software Bareos (see http://www.bareos.org), 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 (http://www.bacula.org).

In addition to software backups, the virtual machines are completely backed up as virtual machine image snapshots via Proxmox vzdump (see https://pve.proxmox.com/wiki/Backup_and_Restore), 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.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

17. Comments/feedback

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

0. N/A: Not Applicable.

Self-assessment statement:

Thank you for the opprotunity to overhaul the application. I'm a bit puzzled by the extended guidance document, because to me it looks like there are questions with similar aspects which makes the application a bit redundant. Thank you very much for your work however!

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments: