Implementation of the Data Seal of Approval

The Data Seal of Approval board hereby confirms that the Trusted Digital repository CLARIN Center BBAW complies with the guidelines version 2014-2017 set by the Data Seal of Approval Board.
The afore-mentioned repository has therefore acquired the Data Seal of Approval of 2013 on May 8, 2015.

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

Yours sincerely,

The Data Seal of Approval Board

Assessment Information

Guidelines Version:	2014-2017 | July 19, 2013
Guidelines Information Booklet:	DSA-booklet_2014-2017.pdf
All Guidelines Documentation:	Documentation
Repository:	CLARIN Center BBAW
Seal Acquiry Date:	May. 08, 2015
For the latest version of the awarded DSA for this repository please visit our website:	http://assessment.datasealofapproval.org/seals/
Previously Acquired Seals:	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. Repository Context

Applicant Entry

Self-assessment statement:

The Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) has a long tradition of corpus-based lexicography and is committed to open access to its primary research data. A partner of the Common Language Resources and Technology Infrastructure (CLARIN), it has established a CLARIN service center in 2013. Goals of this service center include publication and preservation of historical and contemporary German text corpora as well as the lexical resources provided by the Zentrum Sprache (Language Centre) at the BBAW via a data repository.

Since the usage of the CLARIN infrastructure services (e.g. a special PID system, the Component Metadata Infrastructure CMDI) is obligatory, in case a CLARIN center is unable to continue offering its services, it would (to a certain extent) be possible to move its digital assets to another CLARIN center. The handle PID system is hosted at the 'Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen' (GWDG) which is certified according to ISO 9001:2008 (http://www.gwdg.de/fileadmin/inhaltsbilder/Pdf/Presse/pi-012013_24062013.pdf).

see also: http://www.clarin.eu/content/component-metadata

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

1. The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data, and compliance with disciplinary and ethical norms.

Minimum Required Statement of Compliance:

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

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). Thus, our repository provides a quality assessment by which the data consumer can make some judgment about the level of trust or about the reputation of the depositor on the basis of the meta-information about the source institution/organization information associated with any given resource. Our repository does not (and cannot) systematically verify whether the data received have been collected according to these quality standards. 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/legal_conditions/good_scientific_practice/index.html

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:

2. The data producer provides the data in formats recommended by the data repository.

Minimum Required Statement of Compliance:

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

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 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

According to the repository documentation in the workflow image on the right handside, our staff inspects the the depositors files to ensure that the files meet the repositorys 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:

3. The data producer provides the data together with the metadata requested by the data repository.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

CMDI metadata is uploaded or created during the archiving process. This step is required during the uploading process, since data without metadata is technically not accepted by the system. The front-end of the archiving system includes software to assist the depositor in creating valid CMDI metadata.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

4. The data repository has an explicit mission in the area of digital archiving and promulgates it.

Minimum Required Statement of Compliance:

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

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 the repository is to ensure the availability and long-term preservation of german text corpora, lexical and other resources.

This mission is supported by the infrastructure of the Berlin-Brandenburg Academy of Sciences and Humanities and by the integration of the repository into the national and international CLARIN infrastructures.

As part of the CLARIN infrastructure, the repository is included by all promotional activities carried out at the national level of CLARIN-D as well as the European level of CLARIN.

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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

5. The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.

Minimum Required Statement of Compliance:

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

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 licences 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

Example documents can be downloaded here:

CLARIN Deposition License Agreement (DELA)

CLARIN End User License Agreement (EULA)

CLARIN Terms of Service (TOS)

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

6. The data repository applies documented processes and procedures for managing data storage.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

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. The backups are copied to tape storage which is deposited in a locked safe in a separate fire safety zone of the building (in german: 'Brandschutzabschnitt') and are performed with open source software, so that they are recoverable also on a long-term basis.

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 Amanda (see http://www.amanda.org), which determines independently when incremental and full dumps have to be made (but full dumps are done at least once per month). Amanda is open source software which is based on basic GNU backup software such as tar, gzip and dump, which ensures the ability to recover backups even in the distant future. In addition to software backups, the virtual machines are completely backed up as virtual machine image snapshots via Proxmox vzdump (see http://pve.proxmox.com/wiki/Backup_-_Restore_-_Live_Migration), 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 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.

Data integrity is achieved by checking MD5 hash values of the datasets on a regular basis. see

http://clarin.bbaw.de/en/documentation/#Data_Management

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

7. The data repository has a plan for long-term preservation of its digital assets.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

In addition to the measures mentioned under §6 above to ensure the preservation of the raw resource data, 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. Many parts of the CLARIN infrastructure do address the migration of data from one resource center / repository to another. Since the usage of these infrastructure services (e.g. a PID system, CMDI) is obligatory, every CLARIN center is, to a certain extent, ready to move it's digital assets to another center. This is of paramount importance in case a center/repository would be unable to continue offering its services. The virtual machines can be hosted by other centres.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

8. Archiving takes place according to explicit work flows across the data life cycle.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

CLARIN-D has contributed a userguide (http://de.clarin.eu/en/language-resources/userguide.html) which serves as a compehensive overview on the CLARIN-D infrastructure and describes many best practices used at the service centers.

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

- DTA Basisformat text encoding metadata, see: http://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf
- 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: http://www.clarin.eu/cmdi

The ingest, management and storage procedures are described in this workflow chart: See http://clarin.bbaw.de/en/documentation#Workflow and http://clarin.bbaw.de/en/documentation#Quality_management

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 mainly decides what material is being archived; the archive only has technical requirements with regard to the file formats and encodings. 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.

Concerning the presentation side of the data, there is a paper available dealing with extensions to the query languages which were added to deal properly with historic texts. See: http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

9. The data repository assumes responsibility from the data producers for access and availability of the digital objects.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

In general it is the BBAW policy to accept only resources that are available for scientific use (preferably under a Creative Commons License like CC-BY-SA as recommended in the repository documentation). All archived resources are available online, the access permissions are defined by the depositors. Crisis management is addressed on a technical level. Since a PID system is used in CLARIN, moving resources from one CLARIN resource center to another one is possible without affecting the validity of references (e.g. PID reference of a resource used in a research paper). Our setup consists of virtual machines which are implemented by a high-availability failover cluster.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

10. The data repository enables the users to discover and use the data and refer to them in a persistent way.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

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/search/adv_search) is provided, as well as a simple search tool (http://clarin.bbaw.de/search/) for textual content. All metadata can be harvested via the OAI-PMH protocol. Unique persistent identifiers according to the Handle system are provided for each corpus and the each session within the corpora. 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 FCS (http://weblicht.sfs.uni-tuebingen.de/Aggregator/) to enable fulltext search across all text corpora available in all Clarin centers. 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:

11. The data repository ensures the integrity of the digital objects and the metadata.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The integrity of the data is ensured by the version control in the Fedora-Commons back-end by MD5 checksums. Checksum tests are done regularly, especially before performing backups. Metadata is a data stream within the digital object, and as such is version-controlled like object data. The availability of file, web, and application servers is monitored continuously. We consider all objects deposited in our repository as fixed and immutable. We create new digital objects for updates and keep the old versions in our repository. However, updates of metadata for existing resources are possible without considering the result to be a new version.

All previous versions of a newly submitted existing digital object can be reached via links the dropdown object history in the web frontend.

We changed the Statement of compliance to to 'fully implemented' because after some serious work on the software we consider the workflow as stable.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

12. The data repository ensures the authenticity of the digital objects and the metadata.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The repository in principle makes the original deposited objects available in an unmodified way, if the objects are delivered in one of the accepted file types and encodings. New versions of archived resources can be deposited, in which case the old versions will be moved to a version archive. Different versions of the same resource are not compared; we assume the depositor has good reasons for depositing a newer version. A new version of a resource will get a new persistent identifier; the old version will keep the original persistent identifier. Metadata can change if the depositor or archivist sees the need for that, in the case of errors or missing information. Changes to the metadata are currently not logged. All archived objects are linked to their metadata descriptions and are organized in hierarchical (or multi-rooted) tree structures to indicate relationships between objects and sets of objects. The tree structures can change if the depositors decide that this is necessary. The identities of the depositors are checked by the repository staff when they hand over their data. Provenance metadata as to who made changes to the repository is currently only stored in log files and not shown to the data consumer.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

13. The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.

Minimum Required Statement of Compliance:

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

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 rely on the group of emerging standards around CMDI (ISO-CD 24622-1). 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 zipfiles - the packages are virtual (or better: the files are linked to eachother) 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 for content integrity of the data, version management and the connection of data and metadata. Implications for the BBAW CLARIN Center: Content integrity is regurarly checked via MD5 checksums. 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 continously 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 wether 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 http://clarin.bbaw.de , Virtual Language Observatory (VLO) at http://catalog.clarin.eu/vlo , CLARIN Federated Content Search (FCS) at http://weblicht.sfs.uni-tuebingen.de/Aggregator/ and the OAI/PMH-Gateway at http://clarin.bbaw.de:8088/oaiprovider?verb=Identify.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

14. The data consumer complies with access regulations set by the data repository.

Minimum Required Statement of Compliance:

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

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 data in the repository have Creative Commons licenses applied to them. 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.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

15. The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

There are a number of specific codes of conduct that are applicable to parts of the repository, e.g. the DFG ethical Codes of Conduct (e.g. layed down in the DFG Rules of Good Scientific Practice). The codes of conduct are in line with generally accepted codes of conduct for research data in Germany. Any data user is bound by the terms and conditions of use of the repository, as soon as repository services or data deposited in the repository are used. Ethical rules (see also answer for guideline 2):

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/legal_conditions/good_scientific_practice/index.html

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

Generally we won't publish confidential data via the repository.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

16. The data consumer respects the applicable licences of the data repository regarding the use of the data.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

If applicable, the data consumer is made aware of usage restrictions for the data to which she/he has received access. Generally, the usage restrictions are already described in the codes of conduct. For some data, explicit statements need to be made by the data consumer about the use of the data before he/she receives access. The depositor then decides whether or not access is granted. In case of misuse, the only thing that can be done in practice is to deny the user further access to the repository and to make the research community aware of the misuse.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments: