Assessment Information

CoreTrustSeal Requirements 2017–2019

Repository:	BAS CLARIN
Website:	https://clarin.phonetik.uni-muenchen.de/BASRepository/
Certification Date:	17 May 2019

This repository is owned by: **University of Munich**

BAS CLARIN

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository, Research project repository

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository's Designated Community.

As a domain based repository the designated community of the BAS repository are national and international researchers in the fields of human speech and language, speech technology, and speech disorders who work with and create speech resources for empirical research and technology development with a focus on the German language. The BAS also supports national and international researchers in these fields by providing a long-term archive of data and metadata arising from research projects.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

B. Basic curation – e.g. brief checking; addition of basic metadata or documentation, C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

The BAS carries out mainly B and C level curation (A is not supported). At the minimum this entails metadata encoded in CMDI [1] describing the resource, and ensuring availability of data in appropriate formats (see also R7 and R8).

[1] Component Metadata Infrastructure: https://www.clarin.eu/content/component-metadata

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Outsource Partners. If applicable, please list them.

a) Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
The repository makes use of a common CLARIN PID service [1] based on the Handle System [2] and in cooperation with the European Persistent Identifier Consortium (EPIC). The usage of PIDs is mandatory for resources in CLARIN thus all resources added to the repository may be referenced using PIDs. CLARIN-D has a contractual relationship with GWDG concerning the provision of PID-services via EPIC API v2. The following document lists the services which were stipulated [3]

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

CLARIN-D offers several services to it's member institutions, among them the following:
- CLARIN-D HelpDesk (https://support.clarin-d.de/mail/): A central system for user support, which allows for the distribution of user questions and feedback to qualified personnel at the centres.
- CLARIN-D website (https://clarin-d.de/en/): A starting point for researchers to find information on CLARIN-D and to

access CLARIN-D services.
- CLARIN-D wiki (https://www.clarin-d.de/mwiki/index.php/Hauptseite): A central platform for CLARIN-D-related staff.
- CLARIN central monitoring (https://monitoring.clarin.eu/): A monitoring service offered to all CLARIN-ERIC members and maintained by the resource centre Leipzig.

c) CLARIN-ERIC
CLARIN-D is a member of CLARIN'S European Research Infrastructure Consortium (ERIC). CLARIN-ERIC offers central services to it's members and users, as stated in [4]. The services are available to all centres in the member countries of the CLARIN-ERIC [5].

Most important services of the ERIC cover the search functionality for the German CLARIN- centres:

- Virtual Language Obervatory - VLO (https://vlo.clarin.eu): CLARIN's central metadata-based search engine, which contains metadata of all German CLARIN-centres.
- Metadata harvester: The VLO is kept up to date using the metadata harvester run by the CLARIN- ERIC.
- Federated Content Search - FCS (https://www.clarin.eu/contentsearch): Optionally, centres can provide the actual data of their resources for this central content search.

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

d) Leibniz Rechenzentrum Garching
BAS relies on the Leibniz Rechenzentrum Garching (LRZ), operated by the Bavarian Academy of Science, for long term backup and archiving. The corresponding contract with LRZ is renewed every year (see also R9).

[1] https://www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf
[2] http://www.handle.net/
[3] http://www.clarin-d.de/mwiki/images/0/0b/GWDG_PID.pdf
[4] https://www.clarin.eu/value-proposition
[5] https://www.clarin.eu/content/overview-clarin-centres

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

BAS offers its repository services to speech-related projects so that funding agencies' requirements related to long-term availability are met. The number of resources produced by external projects and deposited in the BAS repository is slowly growing. Examples include the Nautilus corpus (TU Berlin and Telekom), WaSeP (Leibniz-Institut für Neurobiologie, Magdeburg), MOCHA (Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh). A number of ongoing projects have requested statements that the BAS will store and make available the resources compiled in these projects.

To our knowledge, BAS is one of the three CLARIN-D centres with a focus on speech, and it is the largest (in terms of hosted corpora and web services) in Germany for phonetics research and technology development. BAS has strong ties with ELRA (European Language Resources Association) and LDC (Linguistic Data Consortium).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

ORGANIZATIONAL INFRASTRUCTURE

I. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The Bavarian Archive for Speech Signals (BAS) is a public institution hosted by the Ludwig-Maximilians-Universität München founded with the aim of making speech resources of contemporary spoken German available to research and speech technology communities via a maximally comprehensive digital speech-signal database. Speech material will be structured in a manner allowing flexible and precise access, with rich annotations, metadata and linguistic-phonetic evaluation forming an integral part of it. [1]

The first aim of BAS will be to satisfy the immediate demand for spoken language data recorded under controlled conditions of the kind required for speech technology development in German. This will include development of new techniques for efficient handling of and access to very large quantities of phonetic data, independent of the location and the nature of the storage. In addition to typical application-oriented corpora this first aim will concentrate on establishing a representative database of contemporary spoken German.

The second goal consists in the long term development of a (more or less) Complete Phonetic Theory (CPT) of spoken German. For this endeavour, the central category will no longer be the speech sound but rather the word as the lexically given unit. The great variability characterizing the pronunciation of words in running speech as opposed to citation form will be systematically documented and related to the communicative information content [2].

References

[1] http://www.bas.uni-muenchen.de/Bas/BasBaseng.html
[2] Tillmann, H. G., Draxler, C., Kotten, K., Schiel, F. (1995). The Phonetic Goals of the New Bavarian Archive for Speech Signals. In Proceedings of the ICPhS (pp. 550-553)

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

II. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Data in the BAS repository consitst of metadata and primary data – metadata in general is freely available. For primary data, restrictions may apply.

1) All CMDI metadata of the BAS repository are provided via Web API and OAI-PMH without access restrictions according to CLARIN-D recommendations.

2) Part of the primary data is also provided without access restrictions (class PUB), without restrictions for academic users (class ACA), and a small part is protected (class RES). For ACA class data, a Shibboleth AAI account of a European university or from the CLARIN IDP is necessary to access the primary data online. To obtain access to class RES data sets, the explicit permission from the depositor is needed. For all BAS repository data, the data consumer needs to agree with a code of conduct ([1], paragraph 5).

Access to the BAS repository is governed by its terms of use (EULA, [1]), which details terms of service, privacy policy, and regulations for data access. End users have to accept these license terms before getting access.

The BAS cannot monitor data usage explicitly, but if abuse of BAS data is reported to the BAS and can be verified by the BAS, end user licenses may be revoked.

References
[1] https://www.phonetik.uni-muenchen.de/Bas/BasTermsOfUsage_eng.pdf

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

III. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

All archived resources are preserved for the long-term, i.e. in perpetuity. Besides the steps to take care of the bit stream preservation of the resources (see R7) some measures are taken to enhance the chance of future interpretability of the data. The number of accepted file formats is limited, to make future conversions to other formats more feasible. As much as possible open (non-proprietary) file formats are used. For textual resources, XML formats are used whenever possible, to make future interpretation of the files possible even if the tool that was used to create them no longer exists. Text is encoded in Unicode to ensure future interpretability.

The technical structure of the BAS repository (e.g. usage of handle PIDs) allows an easy transfer or restoration from backup of the complete data repository including access mechanisms (Web API, OAI-PMH, backup facilities) to another institution. Due to the Archive's strong ties to CLARIN-D such a transfer of the repository to another CLARIN-D institution taking over responsibility is possible any time.

All CLARIN centres commit to ensuring long-term availability, access and to preservation of datasets submitted to their repositories, as set out in their Mission statements. CLARIN centres are set up as a distributed network, where each centre institution is a hub of the digital humanities and brings its own financial resources into CLARIN-D, which ensures continued availability. In this case, the funding by LMU München can at least ensure the intermediate-term maintenance of the infrastructure of center. Additionally, 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 templates of license agreements provided in CLARIN. A memorandum of understanding related to the handover of resources can be found here [1], [2]

[1] https://www.clarin-d.net/de/ueber/zentren/gegenseitige-datenuebernahme (in German)
[2] https://www.clarin-d.net/en/about/centres/mou-taking-other-centre-s-data (in English)

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IV. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Only speech data are to be ingested in the BAS repository that fullfill certain requirements as published in [5]. This ensures that no data with unethical content, produced by non-appropriate procedures or data that form a disclosure risk are ingested in the BAS repository. Speech data and their annotations and their metadata are inspected technically and manually in the compulsory validation and evaluation procedure before ingest (see guidlines published in [6]) – data not conforming the basic requirements or not being validated or evaluated successfully are rejected.

Depositors must sign an agreement stating that they respect IPR (Intellectual Property Rights) and privacy issues and that they own all necessary rights required to deposit the data. In particular, data must be anonymized when applicable and the depositor has to prove that he has obtained proper written usage agreements by the speakers recorded ([1]). The depositor can choose to make the data publicly available (class PUB), restrict access to academics via AAI (class ACA), or to restrict access to individual users (class RES)).

Users of BAS data must confirm that they will use resources only for the intended purpose and in an ethical way ([2]).

Guidelines and model contracts are provided for both, depositors and users on the BAS web pages [3]. Model contracts for Depositors are tailored individually for each depositor.

A recommended template for the declaration of consent of speakers is in [4].

References

[1] http://www.phonetik.uni-muenchen.de/Bas/BasTemplateContract.pdf
[2] https://www.phonetik.uni-muenchen.de/Bas/BasTermsOfUsage_eng.pdf
[3] http://hdl.handle.net/11858/00-1779-0000-000C-DAAF-B
[4] http://www.phonetik.uni-muenchen.de/Bas/BasTemplateInformedConsent_en.pdf
[5] https://www.phonetik.uni-muenchen.de/Bas/BasPolicyExternalResources_eng.pdf
[6] http://www.bas.uni-muenchen.de/forschung/BITS/TP2/Cookbook/

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

V. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The BAS was founded in 1995 and since then it is a part of the Ludwig-Maximilians-Universität München (LMU, Fak. 13 Sprach- und Literaturwissenschaften, Dept. II) and hosted by the Institute of Phonetics and Speech Processing (IPS) of LMU. LMU assigned two permanent positions with the maintenance of the BAS, one for the management of the repository (content acquisition, data management negotiations, CLARIN integration, etc.), the other for the technical development (web services, data migration, etc.).

The IPS has its own IT-infrastructure with a full-time permanent position for administration and operation (for a summary in German, see [1]). Furthermore, the IT-infrastructure is embedded into the Munich research network (Münchner Wissenschaftsnetz MWN) maintained by the Leibniz Rechenzentrum (LRZ), a state-funded computing centre to provide IT- and network services to the Munich universities (see [2], in German).

BAS and IT staff are encouraged to attend training and professional courses, e.g. data management and security workshops as provided by LRZ.

BAS and IT staff are professionals from the respective fields.

By being part of the CLARIN-D consortium the repository also gains access to funding for running and further developing a sustainable repository and resource centre to support these goals. Besides staff resources this includes a budget for attending national and international meetings such as conferences, workshops or internal developer meetings and meetings with the subject-specific working groups.

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. [3] (in German) and [4] (in English) outline the BMBF funding of the centre.

References

[1] http://www.phonetik.uni-muenchen.de/institut/it-dienste/index.html
[2] https://www.lrz.de/wir/regelwerk/geschaeftsordnung/
[3] https://www.clarin-d.net/de/ueber/zentren/arbeitsteilung
[4] https://www.clarin-d.net/en/about/centres/division-of-labour-of-clarin-d-centres

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VI. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

As a CLARIN centre the BAS repository draws on the expert guidance provided by the external advisory committees of CLARIN-D.

CLARIN-D is supported by external advisory committees.

1) 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's-eye view of other developments in the communities. With experienced experts from various backgrounds, a high-profile international committee was formed for this purpose. Members of the IAB are currently: Helen Aristar-Dry, Reinhard Altenhöner, Christiane Fellbaum, Björn Granström, Klaus Schindel, Helge Kahler, Jan Christoph Meister, John Nerbonne, Heike Renner-Westermann and Achim Streit.

2) The joint Technical Advisory Board (TAB) of CLARIN-D and DARIAH-DE, is a committee that supports collaboration on the fundamental technical level between the 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. Members of the Technical Advisory Board are currently: Sabine Roller (University of Siegen, head of the centre of information and media technologies), Karlheinz Mörth (Austrian Acacemy of Sciences), Wolfgang Nagel (Technical University of Dresden, Head of the centre of information services and high performance computing), Peter Leinen (German national library, head of information technology), Jan Haji■ (Prague Institute of Formal and Applied Linguistics, CLARIN Center), Margareta Hellström (ICOS Carbon Portal staff member)

3) The main line of communication of the BAS repository with its Designated Community is via CLARIN. 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.

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 [1]. For communication, mailing lists and wiki contents are maintained.

4) The BAS actively disseminates expertise via tutorials at large international and national conferences, e.g. Interspeech and LREC.

5) The BAS repository takes an active role in CLARIN-D's helpdesk. Help is available to any user of the BAS repository, regardless of a membership in CLARIN.

[1] https://www.clarin-d.net/en/disciplines

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

DIGITAL OBJECT MANAGEMENT

VII. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The repository in principle makes the original deposited objects available in an unmodified way, if the objects are in one of the accepted file types and encodings. In case of changes by the data producer, the repository creates a new digital object with a new PID, which refers to the previous version via its PID. In the case that the repository has to change the data, e.g. because a file format becomes obsolete and superseded, the original data are kept.

All resources in the BAS Repository (metadata and actual data) are equipped with a checksum, which is checked on a regular basis in coordination with the backup schedule described in R9.

The repository only accepts works from the original data producers, who are acknowledged as such by means of elements in the CMDI metadata. We use CMDI relations (depending on the profile) to link between objects within a collection, and providing links from objects to additional information. An example CMDI record for the ALC corpus is available at [1]. For more technical details on the underlying data modelling see R8 and on the ingest processes which generate these interrelated objects see R12.

External deposits are only accepted after a due diligence process involving a check of the identity of depositors and clarification of all legal issues along the lines described in R2 and R4.

References

[1] http://hdl.handle.net/11022/1009-0000-0001-88E5-3

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VIII. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Policies and criteria for depositing spoken corpora, and other German language resources are available in [1].

Following general CLARIN standards, metadata for the BAS Repository must be provided in the CMDI format with unique references to the actual resources. Comprehensive documentation on how to create CMDI compliant metadata profiles and instances is available at [2].

The creation of metadata files (instances) can be performed with any standard XML Editor, e.g. the XML Editor ARBIL [3] that comes with CMDI support. Additionally, a set of tools is provided that allow data producers to create new or adapt existing metadata to the CMDI standard. To support the depositors, BAS provides the web service COALA [4] to generate valid CMDI metadata for speech resources from tabular text data.

Metadata elements must be compliant to the standards for spoken language resources set in CMDI. These standards are

defined in two data registry profiles: media-corpus-profile and media-session-profile [5]. Furthermore, all data deposits are validated prior to being ingested [6].

If data deposits and/or their metadata do not meet the validation criteria for long-term preservation, BAS negotiates with the data providers. If no agreement is achieved, data deposits are rejected.

For other formats we offer advice for conversion. However, as a general principle we also archive digital data additionally in their original format in order to minimize the risk of conversion loss.

References

[1] https://www.phonetik.uni-muenchen.de/Bas/BasPolicyExternalResources_eng.pdf
[2] http://www.clarin.eu/cmdi
[3] https://tla.mpi.nl/tools/tla-tools/arbil/
[4] https://clarin.phonetik.uni-muenchen.de/BASWebServices/#!/services/Coala
[5] http://www.clarin.eu/ccr
[6] http://www.bas.uni-muenchen.de/forschung/BITS/TP2/Cookbook/

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IX. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:

4 – The guideline has been fully implemented in the repository

Response:

BAS repository procedures (e.g. validation, ingest, version update, backup, metadata schema checks) are documented in the BAS internal documentation directory on the server.

The BAS servers use RAID-5 storage systems. Thus, faulty hardware and data restoration can be performed without interrupting the server with only minor performance degradation.

Access to the server data is restricted by a firewall. The BAS repository, that is data and operating system, is backed-up daily with full backups. Backups have a retention period of six months and are stored on a dedicated backup server in an IBM Tivoli system at the Leibniz Rechenzentrum (LRZ). LRZ mirrors the BAS repository data every night to the Kernforschungszentrum in Jülich.

The integrity of the data is ensured by the version control in our repository software.

User access to the data is restricted through access privileges. Write and update privileges are granted only to system administrators and CLARIN developers; permitted users of the repository have read-only privileges (user authentification via AAI). A dedicated data drop-off directory with write-only privileges is provided to allow users to deposit data.

CLARIN propagates the idea of reproducible research. Thus updates/new versions of resources typically are equipped with a new PID. Only marginal changes to CMDI metadata are versioned without registering a new PID.

Part of the archiving workflow is the integrity check of the data and the metadata by the archive manager. The metadata is parsed for syntactic correctness, 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 of the ingest part of the archiving workflow see also R12.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

X. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Reviewer 2

Comments:
3 – The repository is in the implementation phase

Response:

Legal arrangements:

External depositors have to sign a depositor agreement. These contracts contain statements on

1. the involved parties
2. licenses and copyright
3. rights and responsibilities of the depositor and the repository
4. the content to be deposited
5. removal of content and access conditions
6. availability to third parties
7. provisions relating to use by third parties (e.g. conditions, royalties)
8. liability
10. term and termination of the Agreement

The current version (which will be subject to change based on future experience with depositors) is available on the repository website [1].

The data provider retains all intellectual property rights to their data. The depositor must grant distribution rights to the repository. Access is provided by the repository and distribution rights are specified in the written agreement. Enforcing licenses by data users in the case of misuse is conducted by the property rights owner.

Technical preservation arrangements:

• integrity tests of data stored (MD5)
• periodical local and distributed backups (located in dedicated computing centers with strict access control)
• repeated testing of reinstalling the repository from backup
• administrator access to the repository is limited to a small group of trained experts

Crisis management concerning the availability of the digital objects is addressed on a technical level (described in R9).

Since a PID system is used in CLARIN, moving resources from one CLARIN resource center to another one is possible without affecting the validity of references (e.g. PID of a resources used in a paper).

Our setup consists of standard journaled UNIX file systems (ext4) which may be moved to other CLARIN partners. In case file systems are moved internally (inside the CLARIN-D center) this will be possible without severe impact to user experience (live migration is supported). In case the file systems need to be moved to other CLARIN partners or need to be restored from the backup system a limited downtime will occur.

Legal aspects of the process of relocating data to another institution is addressed by templates of license agreements provided in CLARIN.

By encouraging data depositors to use standardized formats (standard media formats, UTF-8, documented XML, ...) we try to minimize the cases in which obsolescence of file formats will occur in the near future. By enforcing a detailed and exhaustive documentation in case proprietary or "custom" formats are used we ensure that exhaustive documentation is available under all circumstances. Thus it will, at least in theory, be possible to specify and implement data converters.

Long term data usability is ensured by the following measures:

1. We make sure that all data formats, also proprietary ones, are well documented.
2. We enforce provision of information on authorship of the data and encourage adding references to scientific papers describing the data and usage scenarios.
3. Access to data and metadata is provided via widely used open source software stacks (Apache, Tomcat). This maximizes the probability of long term support (updates, security fixes) for the tools being used and improves the ability to run installations of these software stacks independent from the underlying hardware/operating system.

For further information please refer to the repository technical description provided on the repository website [2]

[1] http://www.phonetik.uni-muenchen.de/Bas/BasTemplateContract.pdf
[2] http://www.bas.uni-muenchen.de/Bas/BasRepository_eng.pdf

Reviewer Entry
Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XI. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The BAS repository only accepts metadata matching the CMDI profiles media-corpus-profile and media-session-profile, and primary data must be in one of the accepted file formats. These profiles make sure that sufficient information about the data is present in the repository and visible to the user.

Prior to ingest, all metadata and primary data are validated (see R4), and only valid data is accepted. Furthermore, the validation report is part of the corpus documentation (e.g. [1] for the RVG-J corpus). This documentation is accessible upon login via the corpus landing page (see [2] for an example from the MOCHA corpus).

The BAS 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.

References

[1] http://www.bas.uni-muenchen.de/forschung/BITS/Revalidation_RVG-J.html

[2] http://hdl.handle.net/11022/1009-0000-0007-C2B1-5

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XII. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The BAS Repository uses a proprietary repository system based on a file system on the server (see [1], Chapter 1). It consists of the following components:

- a public sector containing the landing pages and the metadata (CMDI files) of stored corpora and corpus sessions. The PHP landing pages are generated dynamically from the CMDI files,
- a limited-access sector containing the primary data. This area can be accessed by authorized users after Shibboleth authentification,
- an OAI-PMH endpoint providing metadata in CMDI and Dublin Core format. This endpoint can be queried to test its status [2]
- a user search SQL database and interface.

The ingest workflows of the BAS Repository are documented in the public document [1], Chapter 3. Data enters the repository via one of two ways: ingest or update.

1) Ingest means that a new corpus is created in the repository. At BAS, ingest is an automatic process: a script retrieves primary and meta data from the local file system, requests PIDs for the appropriate data items. This script is a proprietary perl script, and it relies on a small set of human- and machine-readable configuration files. The corpus and session data receives the version number 1.

2) Update means that existing data in the repository is modified. Updates occur at irregular intervals, in general as the result of error corrections or extensions of an existing corpus. Again, this is an automatic process. The script uses the same configuration files as the ingest script. It retrieves all modified primary and meta data from the local file system and requests new PIDs for the appropriate data items. The version counter of the updated resources is incremented.

References

[1] http://www.bas.uni-muenchen.de/Bas/BasRepository_eng.pdf
[2] http://www.phonetik.uni-muenchen.de/cgi-bin/BASRepository/oaipmh/oai.pl?verb=Identify

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIII. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The BAS 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. For very large resources where online access is not (yet) technically feasible we also provide the possibility to distribute resources on standard media (such as DVD-R and/or hard discs). An advanced metadata search utility is provided, as well as a simple search tool for textual content. All metadata can be harvested via the OAI-PMH protocol. These features are directly accessible from the repository home page, which can also be reached via a PID-handle [1].

The CLARIN Virtual Language Observatory (VLO) [2] harvests the metadata in CMDI format from all CLARIN centres via OAI-PMH. Metadata from all CLARIN centres (and other relevant archives and repositories) are browsable and searchable via the VLO website. CLARIN has defined a set of facets to narrow down the selection of resources in the VLO. These facets are again based on concept sets and allow access to potential heterogeneous metadata stocks. The search in the VLO combines a full text query with a selection of (multiple) values in facets.

Moreover, the BAS repository is also indexed by other registries (e.g. Reuters data index, OLAC, ELDA, dbis, r3data).

For citation purposes, unique persistent identifiers according to the Handle system are provided for each corpus and each session within the corpora.

References

[1] http://hdl.handle.net/11022/1009-0000-0001-231F-6
[2] https://vlo.clarin.eu

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIV. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use

of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The BAS repository closely follows the recommendations for standards and tools for compiling language corpora [1, in German] issued by Deutsche Forschungsgemeinschaft (DFG).

Following [1], the major requirements for accepting resources for long term archival are:

(a) Metadata: Every resource must be provided in a standardized format or an exhaustive documentation of the proprietary format. At least for the entire resource, a minimum set of CMDI descriptor fields as defined in the CMDI media-corpus-profile and media-session-profile must be provided.

(b) Quality Assurance: Only resources that comply with BAS guidelines are considered for deposit. The depositor is required to sign an agreement stating that these guidelines are met (see also R2 and R4).

Data sharing and reuse is promoted by providing access to the data (download, webservices) within the bounds of applicable licenses and free access to metadata (e.g. via the OAI-PMH protocol). The CLARIN infrastructure contains software components such as the VLO [2] which enable users to browse and search through combined catalogs that contain metadata of all CLARIN repositories.

Corpora and web services are migrated to new hardware or formats when necessary (described in [3] section 5), e.g. new versions of CMDI. Data depositors agree to such migrations in their contract with BAS.

References

[1] Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora. http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_

sprachkorpora.pdf

[2] http://www.clarin.eu/vlo/

[3] https://www.phonetik.uni-muenchen.de/Bas/BasPolicyExternalResources_eng.pdf

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

TECHNOLOGY

XV. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The BAS repository relies on a Linux server (Ubuntu v. 14 with long-term support). The BAS Repository uses a proprietary repository system based on the open source OAI-PMH2 XMLFile File-based Data Provider. This repository software is written in Perl and PHP. It requires a web server which is capable to run CGI and PHP scripts, an SQL database, as well as tools for xml validation (xmllint), metadata transformation (xsltproc), and checksum calculation (md5sum) [1].

The repository supports the OAIS reference model's tasks and functions:

Ingest
Prior to ingest, data providers deposit their data and metadata in a storage area outside of the repository. The archive managers validate the data against the list of accepted file formats and encodings, and the metadata against the CMDI schema and profiles. Then, they assign PIDs to corpora and sessions within, and then ingest the data using scripts and configuration files.

Archival Storage
The BAS repository stores its resources on its own RAID-5 compliant server in its own local network protected by a firewall. An automatic incremental backup is performed to the Leibniz Rechenzentrum (LRZ) on a daily basis.

Data Management
The repository uses a custom administration application for data management, including search engine support, access control mechanisms and versioning. Metadata is distributed via the OAI-PMH protocol, supporting selective harvesting as well. Scripts are used to generate statistics and perform consistency checks on a regular basis.

Administration
Using local authentication, authorization and access infrastructure, data managers conduct administrative tasks.

Preservation Planning
To ensure long-term availability, the repository is archived using the IBM TIVOLI archive service of LRZ on special archive nodes that are permanent, i.e. that do not expire (regular archive nodes expire 10 years after the original date of submission). LRZ maintains and updates this archive system. The LRZ archive is mirrored to the Kernforschungszentrum Jülich on a daily basis.

Access
The digital objects are available for reading access via their PID for authorized users, based on the AAI infrastructure of the CLARIN Service Provider Federation and local user management. The PIDs are available in the matadata, which can be harvested via OAI-PMH.

References
[1] http://www.bas.uni-muenchen.de/Bas/BasRepository_eng.pdf

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XVI. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The BAS server is contained in a dedicated climatized server room with restricted physical access (only the system administrators have keys to this room). The server software is updated when updates become available. During the update process, the repository and services are not available. Such planned downtimes are announced at least three days in advance, and the updates are generally carried out on weekends to minimize the impact.

The BAS repository stores its resources on its own RAID-5 compliant server in its own local network protected by a firewall. An automatic full backup is performed to the Leibniz Rechenzentrum (LRZ) on a daily basis using the IBM Tivoli Backup System. The LRZ backup archive is mirrored to the Kernforschungszentrum Jülich on a daily basis.

The local storage and server hardware is replaced at irregular intervals, depending on the technical requirements. For all server hardware we have quick response (6h) maintenance contracts with the suppliers and full guarantee for the life-time (5 years). The server and network infrastructure is maintained by professional system administrator (Dipl.Ing., full-time).

Processes to ingest new corpora, to update metadata information, to update content of corpora including a full versioning system, to move the server location, to maintain and move the web services server, as well as documentation of the used maintenance software are documented in text files in a working space accessible for the CLARIN employees and the system administrator.

Introduction to the LRZ backup storage and guidelines for data recovery can be found in [1] and [2] (in German). BAS follows these guidelines. No further risk management is applied.

References

[1] http://www.lrz.de/services/datenhaltung/adsm/

[2] http://www.lrz.de/services/datenhaltung/adsm/richtlinien/

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

APPLICANT FEEDBACK

Comments/feedback

These requirements are not seen as final, and we value your input to improve the core certification procedure. To this end, please leave any comments you wish to make on both the quality of the Catalogue and its relevance to your organization, as well as any other related thoughts.

Response:

It is quite difficult to determine the focus and context of various requirements, which leads to many responses containing similar or identical text (often copy & paste answers). The questions in the guidance section are more focused. It might thus be a good idea to number these questions and ask the applicant first to provide a short text on the overall requirement, and then to respond to these questions one by one - possibly in a multiple-choice manner (where possible).

Furthermore, all Requirements were pre-set to a minimum statement of compliance at level 0, which surely is not correct. This may be an effect of using the old DSA system. In our response, we have set the compliance level appropriately.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments: