BAS CLARIN

Notes Before Completing the Application

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

True

Reviewer Entry

Reviewer 1

Comments:
Accept

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

Brief Description of Repository

The Bavarian Archive for Speech Signals (BAS) is a public institution hosted by the University of Munich founded with the aim of making speech resources of contemporary spoken German as well as tools for the processing of digitized speech available to research and speech technology communities. 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. Since 20th of June 2013 the BAS is a licensed CLARIN B center.

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).
BAS repostitory contents are reviewed in 5-year cycle to ensure that data formats are still acceptable; if not, contents are transformed loss-less into new appropriate formats (new version).
[1] Component Metadata Infrastructure: https://www.clarin.eu/content/component-metadata

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Insource/Outsource Partners. If applicable, please list them.

In-/Outsource Partners:

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) 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

Summary of Significant Changes Since Last Application (if applicable).

The funding of CLARIN-D has ended; BAS is now a member of NFDI funding 'Text+'.
The raw archive space (spinning disks) increased from 7 TByte to 127 TByte.
The number of archived speech corpora has increased since the last certification by 8 new resources.
The BAS repository is now the official archive for the DFG project 'Oral History Digital' (https://www.oral-history.digital/); this means that historians or institutions who wish to archive their video interview collections can transfer their raw video material plus transcriptions into the BAS respository for long-term archival.

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 two 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).

BAS is (via CLARIN-D) 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.

Since Oct 2020 the BAS repository is official long-term storage for the DFG project 'Oral History Digital' (https://www.oral-history.digital/)

Specific BAS reference data:
re3data.org DOI: 10.17616/R3N59T
PID of home page BAS: hdl.handle.net/11858/00-1779-0000-000C-DAAF-B
PID of BAS repository landing page: http://hdl.handle.net/11858/00-1779-0000-0006-BF00-E

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

ORGANIZATIONAL INFRASTRUCTURE

1. 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
Accept

Reviewer 2

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

Response:

Mission
The Bavarian Archive for Speech Signals (BAS) is a public institution hosted by the University of Munich 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][2]

The BAS Repository is organized according to the FAIR guidelines for scientific data: Findability, Accessibility, Interoperability, Reusability:
Findability (i): Speech resources archived at BAS will be visible in science indices and meta search engines. The metadata (MD) of archived resources follow common MD standards such as DC, OLAC and CMDI, and are exported via a public OAI-PMH interface to harvesting science indices, e.g. Web of Science Data Citation Index, Virtual Language Observatory (https://vlo.clarin.eu/), or Open Language Archives Community (http://www.language-archives.org/). BAS metadata are always public and must not contain any personal data.

Findability (ii): Speech resources archived at BAS can be referenced by a permanent URL called Persistent Identifier (PID). Each corpus (and each recording session within) are assigned to PIDs of the handle system (https://www.handle.net/) that are maintained by the BAS. PIDs allow citations to data sets that will always point to the corresponding repository page or MD record even if the archive has moved to another web address.
Accessibility (i): Speech resources archived at BAS are protected against unauthorized access. The owner of the resource decides according to three basic CLARIN license models who can access her data: ''PUB'' means the resource is available for everyone, ''ACA'' licensed resources are accessible only for academics that can authentificate themselves via the international AAI Shiboleth system, and ''RES'' resources are restricted to a list of users maintained by the owner.
Accessibility (ii): Speech resources archived at BAS are protected against data loss. The BAS Repository is organized according to Open Archival Information System (OAIS) principles (Core Trust Seal, CLARIN Center B evaluations); daily backups via encrypted channels are run to the Leibniz Rechenzentrum Munich and to Rechenzentrum Jülich. Only authorized IT personell at BAS have direct access to archived data.
Accessibility (iii): Speech resources archived at BAS can be accessed any time and without delay by the owner of the data. Data are stored on 'LiveSystem' (spinning discs), not on LTO tapes.
Interoperability: Speech resources archived at BAS are checked periodically for integrity and whether their data formats are still supported by state-of-the-art software. If necessary, archived media streams are transcoded to a new loss-less format.
Reusability: Speech resources archived at BAS can be a valuable source for other disciplines; by means of the BAS repository they are visible in that regard, especially for speech technology applications. [1]

The current BAS Mission has been approved by the host organization, Ludwig-Maximilians-Universitaet Muenchen.
The BAS Mission text in its first version (2000) has been approved by the Scientific Advisory Board
(Paul Dalsgaard (Danmark), Hiroya Fujisaki (Japan), Wolfgang Hess (Germany), Ron Kay (USA),
Joseph Mariani (France) and Wolfgang Wahlster (Germany)).[3]

The mission explicitly refers to the 'Long-term Preservation and Disaster Plan' [4].

[1] Mission: 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)
[3] Advisory Board's Evaluation report 2000: https://www.bas.uni-muenchen.de/Bas/BasEvaluationsbericht2000.pdf (German only)
[4] http://www.phonetik.uni-muenchen.de/Bas/PublicLongtermPreservationPlan.pdf

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

2. 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
Accept

Reviewer 2

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

Response:

Data in the BAS repository consists 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, users can exploit the data scientifically and commercially as stated in the Terms of Usage), 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:

Reviewer 2

Comments:
Accept

3. 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
Accept

Reviewer 2

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

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 interoperability 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 interoperability.
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].

Regarding medium- and long-term plans: the base funding (2 permanent full-time positions) of the host organization LMU is guaranteed until 2029.

[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:

Reviewer 2

Comments:
Accept

4. 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
Accept

Reviewer 2

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

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 guidelines 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/BasTemplateContractEng.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:

Reviewer 2

Comments:
Accept

5. 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
Accept

Reviewer 2

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

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 Text+ consortium (NFDI, [3]) 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. The first (approved) funding phase of Text+ runs until autumn 2026.
Aside from base funding the BAS repository receives funding through several scientific projects, either for further development of the technical infrastructure or for providing repository services (e.g. DFG [4] and BMWi [5])

References

[1] http://www.phonetik.uni-muenchen.de/institut/it-dienste/index.html
[2] https://www.lrz.de/wir/regelwerk/geschaeftsordnung/
[3] https://www.text-plus.org/en/home/#
[4] https://www.oral-history.digital/
[5] https://www.speaker.fraunhofer.de/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

6. 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
Accept

Reviewer 2

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

Response:

Since Oct. 2021, BAS is a member of the German Text+ research data infrastructure initiative ([1]), a successor to CLARIN-D. Text+ distinguishes three Task Areas: Collections, Lexical Resources and Editions, and consists of four institutions as co-applicants:

• Leibniz Institut für Deutsche Sprache Mannheim (IDS, www.ids-mannheim.de),
• Berlin-Brandenburgische Akademie der Wissenschaften (BBAW, www.bbaw.de),
• Deutsche Nationalbibliothek (DNB, www.dnb.de), and
• Staats- und Universitätsbibliothek Göttingen (SUB, www.sub.uni-goettingen.de).

These institutions provide the expertise and technology for the task areas, and they are supported by long-term public funding. Among these institutions, IDS has a strong focus on spoken language, and there is a regular informal exchange between BAS and the spoken language group at IDS.

The governance model of Text+ includes a scientific board with 8 members from the three Task Areas and from Operations; four of these seats are reserved for experts outside of Text+, one for each Task Area plus Operations. Each Task Area features a Scientific Coordination Committee, and there is an Organisational Coordination Committee. These committees review the services and give recommendations to the Scientific Board. An Advisory Board provides feedback to the Scientific Board ([2]).

BAS is a so-called Participant within Text+. A participant receives basic funding from Text+ for the project period, and it may apply to the Scientific Board for additional funding. It reports to the Text+ Scientific Board, which itself reports to the Advisory Board.

References:

[1] www.text-plus.org/en/home/

[2] www.text-plus.org/en/about-us/boards-2/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

DIGITAL OBJECT MANAGEMENT

7. 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
Accept

Reviewer 2

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

Accept

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. Deposits are always supervised by BAS personell; we do not allow automatic deposits (except
by approved project partners).

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. Also 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 in the previous version.
All changes are logged in the version history (not public). See [2] for technical description of these policies.

All resources in the BAS Repository (metadata and actual data) are equipped with a hash/checksum, which is checked on a
regular basis (2 times a year).

The repository only accepts data from the original data producers, who are acknowledged as such by means of elements in the corresponding CMDI metadata. BAS CMDI metadata are always public, and are linked on the data set landing page as well as exported via OAI-PMH. 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
[2] https://www.bas.uni-muenchen.de/Bas/BasRepository_eng.pdf

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

8. 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
Accept

Reviewer 2

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

Response:

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

Primary data (signals and annotations) are only accepted in in standardized, non-proprietary file formats [6]; a list of preferred formats is in [7]

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 CMDI profiles media-corpus-profile and media-session-profile [5]. All metadata (CMDI) are validated automatically before ingest, after version updates, and after metadata updates using the schemata published by the CLARIN
CMDI registry.[6]

If data deposits and/or their metadata do not meet the validation criteria for long-term preservation [6], BAS negotiates with
the data providers. If no agreement is achieved, data deposits are rejected.

For other metadata 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.

Data once ingested are never removed from the repository, so that PIDs are always resolvable. The only exception are data which cannot remain in the repository for ethical reasons or violence of personal data; in this case the PID is re-directed to a dummy page explaining why the data had to be removed.
Ingested data can only be altered by version update; their metadata can be altered without version update.

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] https://www.clarin.eu/componentregistry
[6] http://www.bas.uni-muenchen.de/forschung/BITS/TP2/Cookbook/
[7] http://www.bas.uni-muenchen.de/forschung/Bas/BasFormatseng.html

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

9. 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
Accept

Reviewer 2

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

Response:

BAS repository procedures (e.g. validation, ingest, version update, metadata update, backup, metadata schema checks, integrity checks,
dissemination) are documented in the BAS internal documentation directory on the server, accessible only to system manager and BAS personnel.

The BAS servers use RAID-5 storage systems (no degradable media) and fully redundant RAID controllers, power supplies and network interfaces.
Thus, faulty hardware and data restoration can be performed without interrupting the server with only minor performance degradation.
BAS servers have a full support warranty for their life cycle of 5 years; after 5 years they are replaced.

Internet access to the server data is restricted by two cascaded firewalls, one managed by the Leibniz Rechenzentrum, one by BAS. Server
rooms are only accessible with special transponders.

The BAS repository, that is data and software, is backed-up daily with full change backups. Backups have a retention period of six months
and a depth of 5 last versions of a file, and are stored on a dedicated backup server in an IBM Tivoli system at the Leibniz Rechenzentrum (LRZ)
in Garching north of Munich.
LRZ mirrors the BAS repository data every night to the Kernforschungszentrum in Jülich. The functionality of data recovery is tested within
the regular IT management of the host institution. Apart from the backup locations no mirrors of the repository data exist on our or other servers.

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 Shibboleth). A dedicated data drop-off directory with write-only privileges is provided to allow potential depositors to deposit data
for technical and manual inspection.

CLARIN propagates the idea of reproducible research ('FAIR'). Thus updates/new versions of resources typically are equipped
with a new PID. Only 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 twice a year. The metadata is
parsed for syntactic correctness, completeness and soundness. All data streams of all versions are equipped with a registered
MD5 checksum, which is checked against the archived data twice a year.

For further details of the ingest part of the archiving workflow see also R12.
For more technical details of the server hardware and access security see R15 and R16.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept.

10. Preservation plan

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

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
Accept

Reviewer 2

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

Response:

Public documentation:
All relevant processes, measures and legal information regaring long-term preservation are summarized in the following document published on the BAS web page [6].

In the following we list the main points relevant to R10:

Legal arrangements for data deposits
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. access conditions
6. availability to third parties
7. provisions relating to use by third parties (e.g. conditions, royalties)
8. liability
9. deposit fees
10. term and termination of the Agreement
The current version is available on the repository's website [1].
The depositor always retains all intellectual property rights to their data. The depositor must grant distribution and maintenance rights to the repository. Access as provided by the repository and distribution rights are to be 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
The repository guarantees the integrity and authenticity of the data. 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. Deposits are always supervised by BAS personell; we do not allow automatic deposits (except by approved project partners).
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. Also 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 in the previous version. All changes are logged in the version history. See [2] for the technical implementation of these policies.
All primary resources in the BAS Repository are equipped with a hash/checksum, which is checked on a regular basis.
The repository only accepts data from the original data producers, who are acknowledged as such by means of descriptive elements in the corresponding CMDI metadata. BAS CMDI metadata are always public, and are linked on the data set landing page as well as exported via OAI-PMH. An example CMDI record for the ALC corpus is available at [3].
The following procedures/measures ensure the data integrity of archived data:
• periodical (twice a year) integrity tests of archived digital objects (MD5)

• periodical local and distributed backups (located in dedicated computing centers with strict access control)
• periodical tests of reinstalling the repository from backup
• administrator access to the repository is limited to a small group of trained experts (BAS personnel)
• physical access to servers is restricted to system administrators
• internet access to servers via two cascaded firewalls (one maintained by LRZ, one by BAS)

Long-term data preservation measures

By encouraging data depositors to use standardized formats (standard media formats, UTF-8, documented XML, ...) we minimize the cases in which obsolescence of file formats will occur in the near future. By enforcing a detailed documentation in case proprietary 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, if needed.

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.
4. File formats of archived data are re-evaluated every 5 years for usability; in case that a format is out-of-use, we apply loss-less conversions of the archived data in a contemporary format archived in a new version.
5. There exist no distinction or preference system regarding archived data with regard to preservation; all archived data are treated equally.

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

Disaster plans

In case of (partial) data loss:

The BAS repository, that is data and software, is backed-up daily with full change backups. Backups have a retention period of six months and a depth of 5 last versions of a file, and are stored on a dedicated backup server in an IBM Tivoli system at the Leibniz Rechenzentrum (LRZ)
in Garching north of Munich. LRZ mirrors the BAS repository data every night to the Kernforschungszentrum in Jülich. The functionality of data recovery is tested within the regular IT management of the host institution. Apart from the backup locations no mirrors of the repository data exist on our or other servers.

In case of partial hardware loss (e.g. disc failure, hacks):

The BAS servers use RAID-5 storage systems (no degradable media) and fully redundant RAID controllers, power supplies and network interfaces. Thus, faulty hardware and data restoration can be performed without interrupting the server with only minor performance degradation.

BAS servers have a full support warranty for their life cycle of 5 years; after 5 years they are replaced.

Internet access to the server data is restricted by two cascaded firewalls, one managed by the Leibniz Rechenzentrum,

one by BAS. Server rooms are only accessible with special transponders.
In case of total host server loss:
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. Our technical setup consists of standard journaled UNIX file systems (ext4) which can be moved to other CLARIN partners. In case file systems are moved internally this is 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.
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. 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: [4], [5].
Data ranking:
There is no distinction or preference system regarding archived data with regard to this preservation plan.

References:
[1] http://www.phonetik.uni-muenchen.de/Bas/BasTemplateContractEng.pdf
[2] http://www.bas.uni-muenchen.de/Bas/BasRepository_eng.pdf
[3] http://hdl.handle.net/11022/1009-0000-0001-88E5-3
[4] https://www.clarin-d.net/ueber/zentren/gegenseitige-datenuebernahme (in German)
[5] https://www.clarin-d.net/about/centres/mou-taking-other-centre-s-data (in English)
[6] http://www.bas.uni-muenchen.de/forschung/Bas/PublicLongtermPreservationPlan.pdf

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

11. 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
Accept

Reviewer 2

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

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; the BAS negotiates (and assists) with the
depositor, if rectifications are necessary to meet our validation standard [3]. The
validation report is part of the corpus documentation (e.g. [1] for the PhonDat 1 corpus). It contains the full validation protocol and results,
as well as a validation summary, recommendations for improvement and a quality rating in prose. This documentation is accessible
upon login via the corpus landing page (see [2] for an example landing page of 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] https://www.bas.uni-muenchen.de/forschung/BITS/Revalidation_PD1.html
[2] http://hdl.handle.net/11022/1009-0000-0007-C2B1-5
[3] https://www.bas.uni-muenchen.de/Bas/BITS_Cookbook_TP2.pdf

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

12. 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
Accept

Reviewer 2

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

Response:

The BAS Repository is organized according to Open Archival Information System (OAIS) principles.

Technical setup:
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 of the repo main page, corpus pages and session pages are generated dynamically from the archived CMDI files,
- a limited-access sector containing the protected primary data. This area can be accessed by authorized users after AAI Shibboleth
authentification (academics),
- an OAI-PMH endpoint providing metadata in CMDI and Dublin Core and OLAC format. This endpoint can be queried to test its
status [2]
- a user search SQL database and interface.

Ingest and Updates:
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 but supervised process: a script retrieves
primary and meta data from the local file system, checks and cross-checks them, transfers them into the repository, and 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 of the first ingest
receive the version number 1.
2) Update means that existing data in the repository are 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 (new or updated). The version counter of the updated resources is incremented.

Dissemination:
Dissimination does not require a workflow since the AAI Shibboleth authentication and the access to data is automated. An exception is the
registration of special user access rights (license category RES) which has to be managed in concordance with the copy right holders of the resource.

Deletion:
Ingested data are never removed from the repository; thus PIDs stay always resolvable.

Integrity:
Periodic documented workflows check the integrity of the archived data (twice a year), the correctness of metadata (XML schemata) before ingest/update, and the usability of archived file formats (every 5 years).

Security:
Storage areas and execution rights of automated workflows are setup in a way that only BAS personell can execute workflows regarding the repository. System administrators also have access and execution rights for emergency situations.

Documentation and Change:
All workflows are documented in the internal process documentation accessible only for system administrators and BAS personell.

Workflows can only be changed by common decision of the two permanent BAS employees.

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:

Reviewer 2

Comments:
Accept

13. 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
Accept

Reviewer 2

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

Response:

The BAS repository provides various ways of downloading archived data in formats commonly used by the research communities:
complete corpus, single recording session, single resource file, or a selection of recording sessions provided by the metadata search of the repository

(e.g. "all female speakers aged 35-55 years that are native Italian speakers"). 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 SD cards, USB sticks or SSD).
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) harvests the metadata in CMDI format from all CLARIN centres via OAI-PMH (see [2] for current BAS data in the VLO). 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, re3data).

For citation purposes, unique persistent identifiers using the Handle system are provided for each corpus and for each recording
session within the corpus. On the landing page of each digital object an example citation using the handle PID is shown that end users may utilize.

References

[1] http://hdl.handle.net/11022/1009-0000-0001-231F-6
[2] https://vlo.clarin.eu/search?1&fq=collection:Bavarian+Archive+for+Speech+Signals+(BAS)

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

14. 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
Accept

Reviewer 2

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

Response:

The BAS repository provides the following metadata standards through its web interface and through OAI-PMH export for all archived corpora and recording sessions: CMDI, OLAC, DC.

The BAS repository closely follows the recommendations for standards and tools for compiling language corpora [1, in German] issued by Deutsche Forschungsgemeinschaft (DFG), as well as the recommendations of the CLARIN-D User Guide [4].

Following [1] and [4], the major requirements for accepting resources for long term archival in BAS are [3]:
(a) Metadata: a minimum set of CMDI descriptor fields as defined in the CMDI media-corpus-profile and media-session-profile must be provided, or metadata that can be transformed into the required CMDI metadata.
(b) Every resource must be provided in a standardized format, or - as an exception - an exhaustive documentation of the proprietary format is required.
(c) Quality Assurance: Only resources that comply with BAS guidelines [3] 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) 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.

Archived BAS data are transformed loss-less to new formats when necessary (described in [3] section 5, checks are performed every 5 years).

Data depositors must 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
[4] https://www.clarin-d.net/de/hilfe/benutzerhandbuch

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

TECHNOLOGY

15. 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
Accept

Reviewer 2

Comments:

4 – The guideline has been fully implemented in the repository
Accept

Response:

Hardware:
The BAS repository is installed on a Linux servers (Ubuntu with long-term support) and NAS servers.
The BAS servers use RAID-5 storage systems (no degradable media) with fully redundant RAID controllers, power supplies and network interfaces.
Thus, faulty hardware and data restoration can be performed without interrupting the server with only minor performance degradation.
BAS servers have a full support warranty for their life cycle of 5 years; after 5 years they are replaced.
The raw total capacity is 127 TByte.

Software:
The repository is implemented as a proprietary
repository system written in Perl and PHP following the OAIS model. It requires a web server (Apache) 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].

System security and access:
Internet access is via the DFN implemented by the Leipniz Rechenzentrum Munich with 2 x 1GByte connections
Internet access to the server data is restricted by two cascaded firewalls, one managed by the Leibniz Rechenzentrum, one by BAS. Server
rooms are only accessible with special transponders.

Data security:
The BAS repository, that is data and operating system, is backed-up daily with full change backups. Backups have a retention period of six months
and a depth of 5 last versions of a file, and are stored on a dedicated backup server in an IBM Tivoli system at the Leibniz Rechenzentrum (LRZ)
in Garching north of Munich.
LRZ mirrors the BAS repository data every night to the Kernforschungszentrum in Jülich. The functionality of data recovery is tested within
the regular IT management of the host institution. Apart from the backup locations no mirrors of the repository data exist on our or other servers.
For our disaster plan please refer to R10.

Documentation and maintenance:
System installation, migration and hardware renewal plans are documented in the internal documentation accessible only for system

administrators and BAS personell. The servers are maintained by the system administrator (permanent full position) of the host institution.

References

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

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

16. 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
Accept

Reviewer 2

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

Response:

The BAS servers are located in a dedicated climatized server room with restricted physical access (only the system administrators have access transponders 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 saturdays to minimize the impact.

The BAS repository stores its resources on its own RAID-5 compliant servers in its own local network protected by two firewalls. An automatic full change 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 approximately 5 years 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] (in German).
BAS follows these guidelines. No further risk management is applied.

References

[1] https://doku.lrz.de/display/PUBLIC/Backup+und+Archivierung

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

APPLICANT FEEDBACK

Comments/feedback

These Requirements are not seen as final, and we value your input to improve the CoreTrustSeal certification procedure. Any comments on the quality of the Requirements, their relevance to your organization, or any

other contribution, will be considered as part of future iterations.

Response:

Reviewer Entry

Reviewer 1

Comments:
Although previous source of funding has ended, BAS has demonstrated other source funding and receives funding through several scientific projects, which helps to ensure continuity of services.

Reviewer 2

Comments:
All comments have been addressed. I recommend approving this application.