Implementation of the Data Seal of Approval

The Data Seal of Approval board hereby confirms that the Trusted Digital repository BAS CLARIN complies with the guidelines version 2014-2017 set by the Data Seal of Approval Board.
The afore-mentioned repository has therefore acquired the Data Seal of Approval of 2013 on October 9, 2015.

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

Yours sincerely,

The Data Seal of Approval Board

Assessment Information

Guidelines Version:	2014-2017 | July 19, 2013
Guidelines Information Booklet:	DSA-booklet_2014-2017.pdf
All Guidelines Documentation:	Documentation
Repository:	BAS CLARIN
Seal Acquiry Date:	Oct. 09, 2015
For the latest version of the awarded DSA for this repository please visit our website:	http://assessment.datasealofapproval.org/seals/
Previously Acquired Seals:	Seal date: May 22, 2013 Guidelines version: 2010 | June 1, 2010

This repository is owned by:

Institute of Phonetics and Speech Processing

- Schellingstr. 3
  80799 Munich
  Bavaria
  Germany

  T +49 89 2180 2758
  F +49 89 2180 5790
  E sekretariat@phonetik.uni-muenchen.de
  W http://www.phonetik.uni-muenchen.de/

Assessment

0. Repository Context

Applicant Entry

Self-assessment statement:

1) The document stating the general description of the functions and activities of BAS within CLARIN-D are outlined in the following document:

http://www.bas.uni-muenchen.de/forschung/Bas/BasGeneraleng.html

(last accessed 29.04.2015)

2) List of outsource partners for BAS:

a) Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)

The repository makes use of a common CLARIN PID service (https://www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf) based on the Handle System (http://www.handle.net/) 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 v1 and v2.

b) 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.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The BAS data repository (https://clarin.phonetik.uni-muenchen.de/BASRepository/; last accessed 24.04.2015) contains speech and multimodal corpora for research, education and technology development. The corpora have been created by BAS alone, in collaboration with other academic institutions or industrial partners, or entirely by external partners.

Each corpus contains a plain text description file naming the data producers, outlining the contents, means of data collection, and structure and documentation of the corpus. If corpora are provided by external partners, either these partners also provide this description file, or the BAS creates this description file in collaboration with the external partner.

Each corpus undergoes at least one internal validation. This validation checks the formal consistency of the corpus, but not the content. The validation report becomes part of the corpus documentation and is visible in the repository and corpus web site.

BAS corpus collections underly strict ethical standards: participation is voluntary, participants sign or, in the case of web-based data collection, agree to have read an informed consent form, and where required the data collection procedure must have been accepted by an ethics committee. BAS requests that data produced in collaboration with or by external partners underlie comparable ethical standards, however it cannot systematically verify whether the data it receives is collected according to these rules.

There exist guidelines (in both German and English) for resources created by external partners (https://www.phonetik.uni-muenchen.de/Bas/BasPolicyExternalResources_deu.pdf; https://www.phonetik.uni-muenchen.de/Bas/BasPolicyExternalResources_eng.pdf ; last accessed 28.08.2015)

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

OK, accepted, but the English translation has a wrong URL!

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The data formats used depend on the type of data: signal data in general comes in binary data formats, annotation, documentation and metadata usually comes in text data formats. The BAS requires the use of well-documented and/or publicly specified or de facto standard data formats:

1) binary
1.1) audio: WAV, AIFF, NIST, alaw, mlaw
1.2) video: mpeg2, mp4, QuickTime, AVI
1.3) sensor: device-dependent formats

2) text
2.1) data: plain text (ASCII for legacy corpora, UTF-8 otherwise), XML
2.2) annotation: TextGrid, BPF, EAF
2.3) metadata: CMDI (+ IMDI for legacy corpora)
2.4) documentation: plain text, HTML, PDF

When necessary, BAS performs data conversions from proprietary to open formats, and data migration to accommodate for technology development.

The file formats listed here are fully supported:
http://www.bas.uni-muenchen.de/forschung/Bas/BasFormatseng.html; last accessed 24.04.2015

For the file formats listes here and not in the above document, the BAS will try to offer support:
http://www.clarin.eu/content/standards-and-formats; last accessed 24.04.2015

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

All BAS resources can be accessed and searched via the repository web interface.

BAS corpora in the data repository come with metadata in CMDI and DC format. External partners are requested to provide CMDI metadata along with their corpora. Data whose metadata do not match the profiles are not included in the repository. The BAS offers assistance in creating these metadata where necessary. See the CMDI description (http://www.bas.uni-muenchen.de/Bas/BasSpeechresourceseng.html; last accessed 26.08.2015) for details.

BAS has defined two CMDI profiles (media.corpus.profile and media.session.profile) for speech and multimodal corpora and requires that these profiles be used for the metadata in the repository. These profiles are registered in the CMDI registry and they are on the list of recommended metadata profiles for spoken resources in CLARIN.

To support and guide the creation of metadata for speech resources, there now exists a web service (named COALA) to generate valid metadata for the media-session-profile and media-corpus-profile profiles from Excel or tab-separated text files (http://webapp.phonetik.uni-muenchen.de/BASWebServices/#/services/Coala; click on "show help for this service" on this page; last accessed 24.04.2015).

The technical compliance of the submitted metadata to the CMDI schemas (http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1320657629667/xsd and http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1336550377513/xsd) is validated during ingest.

https://clarin.phonetik.uni-muenchen.de/BASRepository/; last accessed 24.04.2015

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The BAS was founded in 1995 by the Bavarian Ministry of Science and Education and is hosted by the Institute of Phonetics and Speech Processing at LMU Munich. The Institute of Phonetics has charged two staff members with permanent positions (Florian Schiel, Christoph Draxler) with running the BAS.

The main goal of the BAS is to create and make available high quality speech and multimodal corpora for research, education and industrial speech and multimodal technology development.

The BAS has been assigned the official data repository and archive for a number of speech and multimodal corpora in national and academic and/or industrial data collection projects, e.g. Verbmobil, SmartKom and SmartWeb, BITS synthesis and Ph@ttSessionz, and others. Wherever license terms allow it, the BAS has added corpora created during collaboration projects to its catalogue and has continually maintained and updated these corpora so that they remain accessible.

The BAS closely cooperates with the European Language Resources Association ELRA and the Linguistic Data Consortium (LDC), as well as with other resource providers, and it has organized and contributed to workshops and conferences on speech and multimodal corpora.

If funding of the institute is halted, then the CLARIN data and services will be transferred to another CLARIN data centre, preferably in Germany.All URLs used in the repository are PIDs, thus this transfer is transparent to the users.

A document with the mission statement is available here:

http://hdl.handle.net/11858/00-1779-0000-000C-DAAF-B ; last accessed 24.04.2015

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The repository is not a legal entity on its own but is part of the Institute of Phonetics and Speech Processing which is not a legal entity on its own but part of the Ludwig Maximilian University Munich. The legal status of the Ludwig Maximilian University is “Körperschaft des Öffentlichen Rechts“.

The repository is funded by the Institute of Phonetics and Speech Processing. The repository has agreements with external depositors about the right to archive the data. The depositors themselves are responsible for compliance with any legal regulations in the area where the data is collected. The repository enables the depositors to restrict access to their resources at various levels. Distributed copies elsewhere may not be made available to third parties.

Online template contract, code of conduct and terms of usage:

http://www.phonetik.uni-muenchen.de/Bas/BasTemplateContract.pdf
http://www.phonetik.uni-muenchen.de/Bas/BasTemplateInformedConsent_en.pdf
https://www.phonetik.uni-muenchen.de/Bas/BasTermsOfUsage_eng.pdf

(all links were last accessed on 24.04.2015)

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The 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. In addition to the backup, BAS has archived all its corpora using the LRZ IBM TIVOLI archive service 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). The LRZ archive is mirrored to the Kernforschungszentrum Jülich on a daily basis.

Furthermore, a subset of the corpora is also held on optical media in a separate location in the building of the repository.

The local storage hardware is replaced at irregular intervals, depending on the technical requirements.

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.

Introduction to the LRZ backup storage and guidelines for data recovery:
http://www.lrz.de/services/datenhaltung/adsm/
http://www.lrz.de/services/datenhaltung/adsm/richtlinien/

BAS follows these guidelines. No further risk management is applied.

(both links last accessed 24.04.2015)

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

Besides the steps mentioned above for the previous guideline to take care of the bit stream preservation of the resources, 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.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

This is sufficient for compliance statement 3, but it is not really an explicit plan yet.

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

At BAS, the workflow for archiving data consists of the following phases:

1) corpus creation: specification, recording, annotation
2) post processing: formatting, metadata generation, reporting
3) validation: formal consistency checks (completeness, technical validation)
4) distribution and exploitation (repository)

Phase 1) and, in parts phase 2), are outlined in the BAS cookbooks:

http://www.bas.uni-muenchen.de/forschung/BITS/TP1/Cookbook/
http://www.bas.uni-muenchen.de/forschung/BITS/TP2/Cookbook/

Corpora are either created by BAS according to the above workflow, or they are provided by external providers. Data produced externally enters the workflow either at phase 2) or at phase 3).

In phase 4), data enters the repository via one of two ways: ingest or update.

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.

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.

A public description of the workflow is given here:
http://www.bas.uni-muenchen.de/Bas/BasRepository_eng.pdf; last accessed 24.04.2015

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

Information on the "skills of employees" are missing.

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The repository has signed agreements with external depositors. The agreements state the right of the BAS (represented by the Ludwig-Maximilians-Universität München) to archive, maintain and distribute the data to third parties (user licenses); the agreements also state that the purpose of the storage of a resource in the BAS repository is to make the resource available to the scientific community as it is feasible. There is no guarantee that resources are distributed, that is the BAS reserves the right to restrict the distribution for ethical or technical reasons. Access restrictions for certain groups (e.g. commercial enterprises) are defined by the depositors. In general it is the BAS' policy to only accept resources that are available for scientific usage.

The BAS policy for external resources is described in:
http://www.phonetik.uni-muenchen.de/Bas/BasPolicyExternalResources_eng.pdf

A contract template can be found at: http://www.phonetik.uni-muenchen.de/Bas/BasTemplateContract.pdf

(both links were last accessed 24.04.2015)

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The repository provides various ways of utilizing the archived data via online tools as well as by downloading the data in formats commonly used by the research communities. 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. Unique persistent identifiers according to the Handle system are provided for each corpus and each session within the corpora.

Bas repository:
http://clarin.phonetik.uni-muenchen.de/BASRepository/index.php

BAS OAI-PMH endpoint:
https://webapp.phonetik.uni-muenchen.de/BASSRU/
sample query:
https://webapp.phonetik.uni-muenchen.de/BASSRU/?version=1.2&operation=searchRetrieve&query=Gott

(all links were last accessed 24.04.2015)

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

OK, but the presentation is not very user-friendly.

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The repository displays the latest version of all resources. Internally, all resources are governed by a versioning system.

MD5 checksums are calculated for all objects and checked periodically. The availability of files on the file system is checked automatically daily. The availability of the archive access tools is checked automatically several times a day. The availability of file, web and application servers is monitored continuously.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The repository in principle makes the original deposited objects available in an unmodified way, if the objects were in one of the accepted file types and encodings. Additionally, lower quality distribution copies of audio and video recordings may be made available. New versions of archived resources can be deposited, in which case the old versions will be moved to a version archive. The depositor is informed about this by email.

Different versions of the same resource are not compared; we assume the depositor has good reasons for depositing a newer version. A new version of a resource will get a new persistent identifier; the old version will keep the original persistent identifier.

Metadata can change if the depositor or archivist sees the need for that, in the case of errors or missing information. All archived objects are linked to their metadata descriptions and are organized in hierarchical (or multi-rooted) tree structures to indicate relationships between objects and sets of objects. The tree structures can change if the depositors decide that this is necessary.

Changes to the metadata are currently not logged. Such changes are only performed to correct errors in the metadata, and they are performed by staff of the repository. An internal document describes the process, which involves formal checks of the metadata against the schema. However, it is not logged who performed the changes.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

Maintaining provenance and related audit trails would be recommendable for further implementation!

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The repository aims to support 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 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, includin 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. The hardware is securely stored in locations with restricted access.

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.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

Good and clear overview

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

Resource data in the repository is protected, while metadata are openly accessible; an account is necessary to get access to the content data. For some data sets, explicit permission (license) from the depositor is needed; the license has to be filed at the BAS and the data consumer must have a AAI federation user account. For a large part of the data, the data consumer needs to agree with a code of conduct, which also contains licensing terms. If the data consumer does not comply with the access regulations, the only thing that can be practically done is to deny him/her further access and to make the research community aware of the misuse.

http://www.phonetik.uni-muenchen.de/Bas/BasTemplateContract.pdf
http://www.phonetik.uni-muenchen.de/Bas/BasTemplateInformedConsent_en.pdf
http://www.phonetik.uni-muenchen.de/Bas/BasTermsOfUsage_eng.pdf

(all links where last accessed 24.04.2015)

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The repository needs to deal with the codes of conduct relating to processing personal data, e.g. voice data. It does not need to deal with codes of conduct pertaining to protect human subjects. All data consumers must accept the terms of usage as defined in https://www.phonetik.uni-muenchen.de/Bas/BasTermsOfUsage_eng.pdf (last checked 24.04.2015).

No institutional bodies are involved.

Users of the repository will only be granted the requested access credentials if they are a) members of institutions that have agreed to the codes of conduct, or b) if the users have agreed to the codes of conduct before they get access to the data.

All data in the repository are anonymised regarding speaker identity, and no personal confidential information is stored in the repository.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

OK

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

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

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

http://www.phonetik.uni-muenchen.de/Bas/BasTermsOfUsage_eng.pdf (last accessed 24.04.2015)

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments: