Bavarian Archive for Speech Signals (BAS) - Long-term Preservation Plan

F. Schiel - 2022-06-07

This public document documents the processes, measures and legal arrangements that the BAS Repository enforces to ensure long-term preservation of the deposited data.

1. 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.

2. 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 personnel; 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 Also in the case that the repository has to change the data, e.g. because a file format becomes obsolete and superseded, the original data are kept in the previous version.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)

3. Long-term data preservation

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].

4. Disaster plans

<u>In case of data loss:</u>

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.

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

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.

<u>In case of total host server loss:</u>

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)