science & innovation

Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

SADiLaR
South African Centre for Digital Language Resources

Preservation policy
SADiLaR

1. PURPOSE OF THIS DOCUMENT

This document outlines the policies of SADiLaR regarding the long-term preservation of resources. If you have any questions, please feel free to contact us.

2. DATA GENERATION

Data is generated by and assigned a license by the contributor. The contributor also generates metadata in the CMDI format at the time of submission.

For datasets generated through SADiLaR funding, a high standard of data quality for metadata is enforced. For datasets generated outside SADiLaR funding, the metadata quality is the choice of the contributor – however they are encouraged to be as comprehensive as possible and generate high quality metadata.

3. DATA PRESERVATION

One of the fundamental missions of SADiLaR is to deliver long-term access to research data and research tools as well as making it accessible and discoverable. The assets registered within the SADiLaR repository are expected to be persistent and be available indefinitely.

Funding of the SADiLaR is through the Department of Science and Industry

SADiLaR encourages depositors to provide data in CLARIN standard recommended formats, and those used by other repositories, such as the Max Planck Institute for Psycholinguistics' list of accepted formats. Due to the scarcity of language resources, and scarcity of repositories to share these resources, format recommendations are not strictly enforced, but strongly encouraged. In exceptional cases, SADiLaR will assist depositors in converting data to standardised formats where possible.

All data deposited is reviewed annually to determine whether the format of the underlying resource is still widely supported, and in cases where they are not, or may become redundant in the near future, the original depositor is contacted to update the data to a newer format. In some cases, updates to resources may be undertaken by SADiLaR.

Data assets held in the SADiLaR repository are subject to a process of digital migration to support the long-term preservation.

4. DIGITAL MIGRATION

Digital Migration refers to the transfer of digital information with the long-term goal to support data preservation

The process is invoked because of events such as

- Risk of media decay
- Deployment of more cost-effective physical infrastructure
- Data format obsolescence

4.1. Data Refreshment

Data refreshment means that data is moved in physical storage media storage

Such scenario can happen as part of hardware replacement, environmental consolidation, file system optimisation, machine virtualisation.

The data is unchanged, PID unchanged, and Checksum unchanged. The data is still accessible through the PID, still described by the same metadata. All such changes are abstracted from the user community by machine level management.

The Technical Manager, on a continuous operational basis, manages such changes to the environment. This is done in collaboration with service providers delivering IT services

There is periodical review of the underlying technical infrastructure in terms of the factors which drive the need for machine level change such as

- Age of machine; Support period of OS; Overall system capacity in relation to demand; Cost of operation
- Automated machine operation to restructure memory organisation
- VM Migration; Change of hosting environment

Decisions on data refreshment will be taken unilaterally by the technical team that will not affect the delivered service, although downtime maybe needed for some operation.

4.2. Digital Transform

Digital transform means that the data file format is changed with the information encoded unchanged

Digital transform to change the data format, but not the encoded information, is an option when a data format has become deprecated or obsolete or to use the data in a preferred tool. Digital transform is the set of actions to convert the data format to a different data format.

The Technical Manager has the recommended data format on long-term review and will advise whether formats are deprecated or obsolete. Deprecation or obsolescence of a data format from current support does not impact those artefacts submitted historically in that format.

The baseline policy is that data artefacts have no planned digital transform by SADiLaR. Digital transform is the responsibility of the original contributor, although on a case-by-case basis digital transform can be planned and executed by SADiLaR based on demand from the user community. In the case of digital transform the new data artefacts will be submitted to the repository.

In the case of audio, once audio data is encoded by passing through a codec, information can be lost and this is irrecoverable without going back to a higher resolution version of the source data. For example, the popular MP3 and AAC formats lost information when the encoding happened. This cannot be recovered by data migration to a higher resolution format, without using the source data of a higher resolution.

Data contributors are encouraged to encode audio data in lossless format such as WAV.

4.3. Data Change

Data change means that the data in the artefact is to be changed including scenarios of error correction, enhancement and simplification or generation of a derivate work

The baseline policy is that the once data artefacts are onboarded into the repository, they cannot be changed. All data changes will result in a new artefact submission to the repository. This is because the data assets are already open and maybe cited by other researchers.

Data correctness at the time of submission is a core criterion in the submission process and is the responsibility of the individual researcher making the data contribution to the SADiLaR repository. Data errors discovered after archiving to the SADiLaR repo cannot be corrected in the existing artefact as the data is already open to the whole user community and can be in-use by others for creation of derivative work.

Re-issued, corrected or otherwise changed data artefact co-exist with the original artefact in the repository.

4.3.1. Meta-data Updates

Metadata is correct at the time of submission as a core criterion. SADiLaR makes basic checks that the meta-data fields are populated. The individual researcher making the data contribution to the SADiLaR repository is responsible for the meta-data content.

Only the contact details of the contributor in the meta-data can be changed in a maintenance action.

4.4. Data Management Handover

Data management handover means that management of data is handed over between repositories

Data management handover from SADiLaR to another repository would be required should lack of funding of SADiLaR occur, or strategic decision to consolidate part or all of the collection with another collection.

In the case of the closure of SADiLaR, the policy is that the data is handed-over to another repository with all the data, meta-data and PID unchanged. The resource would continue to be permanently accessible despite re-hosting in a different archive. The new archive would have to take over maintenance of the PID. The URI would change but be abstracted from the user community by mapping to the PID

This will be managed on a case-by-case basis. It is expected the receiving repository can download the artefacts and meta-data and from the SADiLAR repository and import into their repository system. PIDs would be repointed to the new host repository.

Rehosting of applications hosted by SADiLaR into other environment will be considered on a case by case basis.

The scope of data management handover excludes user data, e.g. user names and passphrases. Existing users would have to be registered to the new repository.

4.5. Data Destruction

The fundamental value of SADiLaR is that all data and meta-data is permanently available.

In the event of the closure of SADiLaR, all data, associated metadata and PIDs would be transferred to another repository as in 4.4 and once completed; SADiLaR could close the SADiLaR repository including destruction of the data hosted there. Permanency of the data and meta-data will be ensured by the destination repository.

5. VERSIONING

The repository is a permanent archiving system and does not support versioning of changed submissions. All scenarios of data change in 4.3 will result in a new asset in the archive.

6. DATA REFERENCES

All data resource will be allocated a PID (Persistent Identifier) to be used as the URI of the asset. The PID will be allocated indefinitely and will remain valid if the data is moved to a different repository. SADiLaR also makes use of PID's for all deposited resources through the Handle System (https://handle.net).

7. DATA STORAGE AND BACKUPS

SADiLaR is fully supported by the institutional infrastructure of its hosting entity, NWU, for its high availability storage, backup and disaster recovery procedures. Digital storage facilities for all resources are housed in modern server rooms with the necessary cooling, power and fire safety infrastructure, as well as redundancies spread across three remote campuses. Digital access to the storage facilities and virtual machines is protected by firewall, and is limited to a small number of approved SADiLaR employees, and support staff from institutional IT, responsible for the resource availability and backup procedures. Physical access to the servers and storage facilities is restricted to authorised staff by access control systems.

The backup and disaster recovery facilities of SADiLaR resources are provided by institutional IT, in line with the requirements of the NWU. This includes a fully documented and tested backup schedule, consisting of

- incremental daily back-ups, stored for 1 week;
- full weekly backups, stored for four weeks;
- and monthly tape backups, stored for 1 year.

This backup schedule allows backups to be restored, at various point in time, for up to one year. Tape backups are distributed across three remote sites to minimise the possibility of data loss. Most data recovery can be completed within 1 business day, with some more time typically required for tape backup restores.

8. SOFTWARE STACK PRESERVATION

SADiLaR primarily makes use of widely used and supported open-source software stacks for the storage and preservation of resources. This in turn maximises the long term support for the tools and software stacks used in SADiLaR. All digital access is provided via virtual machines following similar backup and data recovery procedures as resource. Security update reviews are performed regularly to ensure that bug-fixes and major updates to software stacks are performed as necessary.