DKRZ Long Term Archive (LTA)
– Preservation and Storage Policy –

version v2.0; 13 December 2022

How to cite:
DKRZ Long Term Archive (LTA) (2022, December 13), Preservation and Storage Policy, Version v2.0, Long Term Archive (LTA) at the German Computing Center (Deutsches Klimarechenzentrum, DKRZ). https://www.wdc-climate.de/docs/DKRZ-LTA-PreservationAndStoragePolicy.pdf (last accessed: YYYY-MM-DD).

These guidelines address the effective implementation of procedures for the preservation of digital data in the Long Term Archive (LTA) of the German Computing Center (Deutsches Klimarechenzentrum, DKRZ), Hamburg, Germany. The long term archiving services provided by the DKRZ LTA comprise the WDCC ("World Data Center for Climate") and DOKU ("DOKUmentation").

Preservation Purpose and Guiding Principles
The main purpose of the DKRZ LTA is to preserve data relevant for climate science and make them available for reuse. The preserved data foremost include output from numerical climate models whereby the focus is on post processed data and not on raw model output.
Data preservation in the WDCC is guided by the FAIR principles.
The WDCC has the data preserved in file formats that are open and well established in the climate science community.
This policy is driven by customer acknowledged guidelines like “Rules of good scientific practice” from the Max Planck Society or the DFG, by community driven standards (like OAIS reference model), Data Seal of Approval (DSA) and other accepted standards like ISO16363 (audit and certification of trustworthy digital repositories).

Storage duration
Data in DKRZ LTA is stored for a minimum of 10 years. After the expiration of this term DKRZ LTA will not remove these data from its archive. However, regular data curation will be stopped, i.e., regular validation of access and integrity (see below) will terminate, unless the data owner opts for a new storage period.

Supported File Formats
Only open source file formats are accepted for preservation. Should individual data objects be stored in a proprietary format, this is only accepted if they are additionally archived in an open source format.
The DKRZ LTA only supports open source formats for preservation that are well established in the climate science community. These comprise:

- **Network Common Data Format (NetCDF)**
  NetCDF is a binary, non-proprietary file format that is self-describing. This means that the metadata describing the data can be stored in the NetCDF file itself. More information on NetCDF can be found at http://www.unidata.ucar.edu/netcdf
  DKRZ LTA strongly recommends to use the Climate and Forecast (CF) metadata convention to standardize the NetCDF data. More information on CF can be found on http://cfconventions.org.

- **GRIdded Binary (GRIB)**
  GRIB is a bit-oriented data exchange and storage format approved by the World Meteorological Organization (WMO). More information can be found on http://www.wmo.int/pages/prog/www/WDM/Guides/Guide-binary-2.html .

Data Integrity
Upon delivery of data, DKRZ staff accepts data objects and its linked metadata, checks its completeness and compliance with community and preservation standards, helps data producers to fix errors or deviations from standards if required, runs checksums and compares them to attached checksums if possible. Only open source data formats are accepted for preservation. Should individual data objects be stored in proprietary formats, those objects are not accepted for preservation unless they are archived in an open source format as well. Metadata, containing provenance information, is compiled from relevant resources and kept in a SQL database. The maintenance, including backup, of this database follows recommended practices by this ensuring metadata integrity and access. On a scheduled basis access and integrity of data objects is validated. In case of errors all necessary steps will be taken to re-enable access to the data objects in its original form.

Data Access
Data objects are made accessible via persistent URLs to the catalog entries. Additionally, WDCC assigned Digital Object Identifiers (DOI) for data collections and Persistent Identifiers (PID) for datasets. The access to data objects is regulated. Only DKRZ LTA staff are authorized to upload data into the archive. Download is possible for registered users via password authentication; restrictions can be implemented if required in consultation with the data owner.

Curation
DKRZ LTA is continuously involved in diverse scientific and technological developments. By this DKRZ LTA staff is in a position to monitor development of community needs and evolution of data formats. Should the need for changes in DKRZ LTA come up, necessary steps are taken. No modification of the data themselves will be performed by DKRZ LTA or the data owner after the data is archived. Should there be any need for changes, modified data objects can be archived as new versions while keeping the old ones.

Storage Infrastructure
DKRZ LTA uses the DKRZ storage infrastructure for its data holdings. In compliance with quality control and security specifications this DKRZ infrastructure is specifically designed to accommodate scalability, reliability and sustainability. Data in the DKRZ Long Term Data Archive is kept in a HSM system at least on two tape copies, internal and external. DKRZ staff actively monitors user requirements and technologies, by adapting DKRZ LTA preservation practices. For issues about security and for a risk assessment see the document DKRZ LTA Risk Assessment which can be found, e.g., on docs.wdc-climate.de. For all matters concerning general protection of personal data see the DKRZ website at https://www.dkrz.de/about-en/contact/en-datenschutzhinweise. This not only refers to data access by web browser but also by any API and by the jblob download tool.

Contacts
In case any questions regarding this preservation and storage policy arise, please contact DKRZ LTA user support at data@dkrz.de.