Preservation Policy

1. Purpose

Digital preservation refers to the series of managed activities necessary to ensure continued access to digital materials beyond the limits of media failure or technological change for as long as necessary (Digital Preservation Handbook, p.24).

This preservation policy describes the digital preservation strategies and principles of the 4TU.ResearchData archive and provides transparency on the procedures involved in ensuring adequate preservation of and access to the data within the archive.

As a trusted digital repository that has received the Data Seal of Approval, 4TU.ResearchData demonstrates its long-standing good archival practice by drafting this preservation policy.

The policy is addressed to data producers and users, as well as research funders, 4TU.ResearchData staff, and other interested audiences.

2. File formats and preservation

One of the goals of digital preservation is to prevent loss of access to files due to file format obsolescence.

To enhance the chance of future interpretability of the data, 4TU.ResearchData has selected a number of preferred formats for full preservation for which it guarantees long-term usability.

In general, the preferred file formats used for full preservation of data, are non-proprietary, well documented, and well understood by 4TU.ResearchData staff.

Data depositors are strongly recommended to provide their data in the preferred format, which is most suitable to the type of data. The 4TU.ResearchData archive also accepts other formats, but informs depositors that only bit-level preservation will be provided.

4TU.ResearchData applies two levels of file support:

- Full preservation: All reasonable actions to maintain usability will be taken. Actions may include migration, normalization or conversion.
- Bit-level preservation: Only access to the object in its submission file format is provided.

The list of the preferred formats is available on our website.

The current accepted and preferred file formats are reviewed periodically to see whether they are still suitable for preservation. The list of preferred formats is updated when necessary.

3. Archival workflow

The 4TU.ResearchData preservation policy adheres to the terminology and preservation practices outlined by the Open Archival Information System (OAIS) Reference Model. The OAIS model provides both a functional model – the specific tasks performed by the archive, such as storage or access – and a corresponding information model, to support long-term maintenance and access to digital material.

An essential element of the OAIS model is the grouping of information into packages:
- SIP: submission information package
- AIP: archival information package
- DIP: dissemination information package
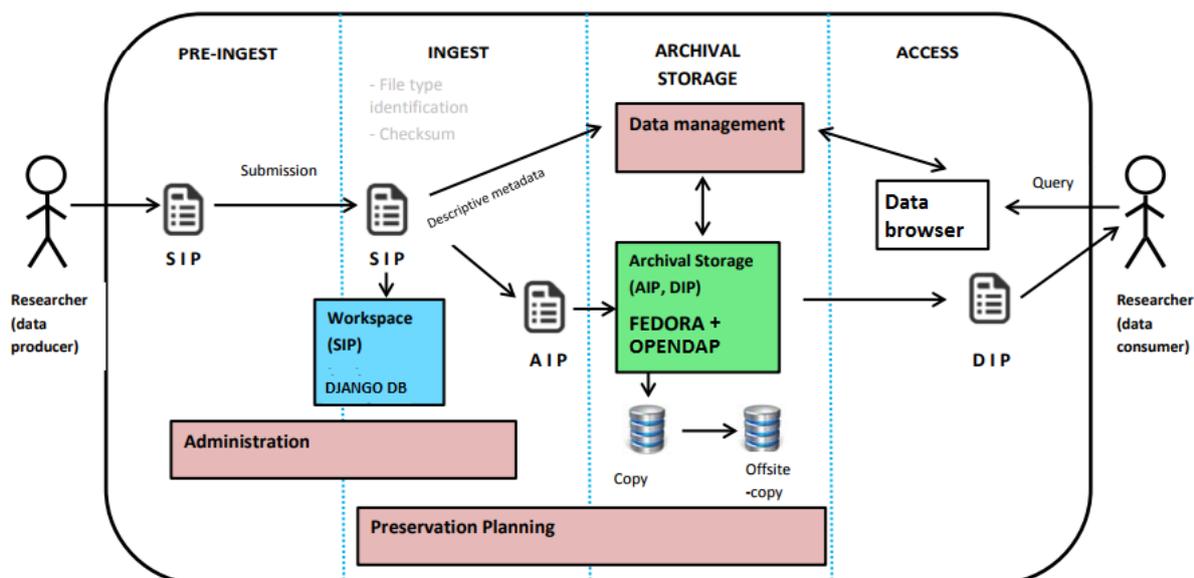
Pre-ingest: Data producers wishing to deposit data are offered guidance and assistance. This can include technical support, but also consultancy about issues such as data formats, the preferred data model, and legal aspects.
Data depositors are provided with clear instructions on how to prepare, document and deposit their data.

Ingest: 4TU.ResearchData collects data from all fields and subjects in science and engineering as outlined in the Data Collection Policy.
Once a data depositor has completed the upload, the data are submitted as a Submission Information Package (SIP) to the data archive. To ensure the completeness, accuracy and usability of the metadata, all deposited datasets undergo the following metadata quality checks: completeness of required and optional metadata, documentation, links to related materials, file format, privacy issues.
Once the SIP is accepted into the system, an Archival Information Package (AIP) is created.
The AIP is the information package the system stores, preserves and sustains.
To ensure the integrity of the datasets, for every deposited file a checksum (md5 type) is made which allows 4TU.ResearchData staff to check the files for defects and restore correct versions if necessary.

Archival storage: The AIP is saved in our Fedora repository (http://fedorarepository.org/). All metadata and data files are saved in Fedora, with the exception of data in netCDF format (https://www.unidata.ucar.edu/software/netcdf/). These files are saved on our OPeNDAP server (https://www.opendap.org/).

4TU.ResearchData uses the DataCite service from the TU Delft Library to assign Digital Object Identifiers (DOIs) to ensure the accessibility and authenticity of the data. Persistent identifiers are minted when items are deposited, and are included with a suggested citation.

When a new version of a dataset is published, a new metadata landing page is created and a new DOI is minted. This way, the already existing persistent identifier (DOI) will continue to refer uniquely to the earlier version of the dataset. The new and the previous dataset are cross-referenced in their respective descriptive metadata.

Access: All datasets are findable and accessible through the Data Browser (https://data.4tu.nl).
The data and their associated metadata and documentation are made available to the useras a Dissemination Information Package (DIP). The DIP is the information package created to distribute the digital content. Users are able to download the complete DIP, as well as separate files within the package.
Access to netCDF data (and HDF5) is further enhanced through the OPeNDAP protocol. A major advantage of using OPeNDAP is the ability to retrieve subsets of files without the need to download the whole dataset, and also the ability to aggregate series of data files, e.g. a time series, into one 'virtual' dataset.

4. Specific preservation actions

Each digital object placed within 4TU.ResearchData is subject to established preservation techniques in order to maintain its integrity and ensure its accessibility over time.

The following actions are undertaken within the preservation process:

Preservation of metadata: Each dataset may consist of several files with different file extensions.
These extensions are mapped to mime types at ingest. The rationale is that (1) datasets may consist of thousands of files, (2) within a dataset, the extensions will have a well-established meaning, but not necessarily between datasets.

The mime types of the files in a dataset are indexed in a dedicated field to allow specific preservation actions to be applied to sets of files of the same type. Over time, multiple versions of a dataset may be created as formats become obsolete. All versions are kept and their creation date are recorded in the metadata. In addition, the bibliographic metadata is augmented with relational metadata (many datasets have semantic relations with other digital objects such as other datasets or representations of instruments). This also facilitates the creation of collections of datasets for a specific preservation purpose.

File format recognition: At the moment, mime types are determined by a mapping from extensions with the possibility to override the standard mappings manually for a specific dataset. We are working towards automatic mime type detection by direct inspection of the files, perhaps augmented with an intelligent guess based on the extension if the inspection gives ambiguous results. The possibility of a manual override will be always maintained.

Secure storage and backup: Data storage of 4TU.ResearchData is managed by the ICT department of Delft University of Technology according to their procedures. The stored research data are backed up (and stored) on hard disks (RAID6) and synchronized (one way) daily. Two times a month a backup is made on disks at another location and retained for one year. Data are stored on two mirror sites.

In order to ensure restore procedures, the root-filesystems are backed up incrementally on a daily basis, and once a week full backups are made. These backups are saved on tapes and will be kept for three months on another location.
A restore can be carried out upon request. Security updates and patches are installed on a regular basis. These preservation procedures are outsourced to the ICT department of Delft University of Technology and recorded in a service level agreement.

Fixity: Fixity means that the digital object has not been changed between two points in time or events. To ensure the integrity of the datasets, for every deposited file a checksum (md5 type) is made which allows 4TU.ResearchData staff to check the files for defects. In case file degradation is discovered, the corrupted data will be removed and replaced with its uncorrupted counterpart from mirror sites.
All changes are logged in an audit trail.

Hardware migration: This means transferring, or rewriting data from an out-of-date medium to a current medium, or transferring data between storage types or computer systems. The purpose of hardware migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology.

File format migration: Changing data from one format to another to ensure the readability and usability of the content. If circumstances dictate that data formats within 4TU.ResearchData are at risk of obsolescence, the content will undergo transformation to a new file format more conducive to its preservation. This may include upgrading datasets to a newer version of the same format or transformation into a completely new file structure. Format migration events will be recorded in the preservation metadata associated with the dataset.

5. Financial sustainability
Long-term sustainability of 4TU.ResearchData requires adequate and reliable sources of funding so that data is preserved properly.
To ensure the data archive is able to fulfil its mission, structural funding is received from the partners[1] of the 4TU.Centre for Research Data consortium.
Should a situation arise which threatens the continued existence of the data archive, these organisations are committed to taking responsibility for the future availability of the data entrusted to the archive.

6. Responsibilities
All 4TU.ResearchData staff work together to ensure that the data stored remains accessible and understandable over the long term. They closely cooperate with the TU Delft ICT department, which is responsible for providing the technical infrastructure.

4TU.ResearchData is responsible for the maintenance, review and revision of all its policies and documentation, including the Preservation Policy.
If you have any comments or questions regarding this policy, please contact us at researchdata@4tu.nl.

7. Related documents
- Data Collection Policy:
  http://researchdata.4tu.nl/fileadmin/editor_upload/pdf/Collection_Policy/Data_collection_policy_aug.2017PDF.pdf
- Data Seal of Approval:
  https://assessment.datasealofapproval.org/assessment_187/seal/pdf/

[1] Currently: TU Delft, Eindhoven University of Technology, University of Twente

- Deposit guidelines:
  http://researchdata.4tu.nl/fileadmin/editor_upload/pdf/Deposit_Guidelines/Deposit_Guidelines.pdf
- Preferred formats:
  http://researchdata.4tu.nl/fileadmin/editor_upload/pdf/File_formats/preffered_file_formats.pdf

8. References
   - Reference Model for an Open Archival Information System (OAIS):
     https://public.ccsds.org/pubs/650x0m2.pdf
   - Preservation Policy DARIS:
     http://forscenter.ch/wp-content/uploads/2015/05/Preservation-Policy_E_v3.pdf
   - Preservation Policy Data Archiving and Networked Services (DANS):
     https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreservationpolicyUK.pdf