4TU. CENTRE FOR RESEARCH DATA

Preservation plan

File formats and preservation

One of the goals of digital preservation is to prevent a loss of access to files due to file format obsolescence.
When a digital object is deposited into a digital repository, the type of file that it is will be declared by its mime type (image/jpeg, application/pdf, etc.). The type of file you are dealing with has big implications for how preservation practices can be applied to it now and in the future. This is because being able to access the contents of a digital object depends on the ability to store, read, and edit the files – actions that are products of the file format's specifications and the software that's necessary to understand that file format.

4TU.Centre for Research Data will take appropriate measures to enhance the chance of future interpretability of the data.
The number of accepted file formats for support levels 1 and 2 (see below) is limited, to make future conversions to other formats more feasible.
In general, the optimal file formats used for long-term preservation of data (or preferred formats), are non-proprietary,well documented, and well understood by 4TU.ResearchData staff.

4TU.ResearchData has adopted hardware migration and file format migration as its primary preservation strategies.
Hardware migration: transferring, or rewriting data from an out-of-date medium to a current medium, or transferring data between storage types or computer systems.
File format migration: changing data from one format to another to ensure the readability and usability of the content.

Levels of file support

4TU.ResearchData applies three levels of file support:
Level 1: All reasonable actions to maintain usability will be taken. Actions may include migration, normalization or conversion.
Level 2: Limited steps to maintain usability will be taken. May actively transform a file from one format to another to mitigate format obsolescence.
Level 3: Only access to the object in its submission file format is provided.

For a complete overview of all file formats and their support level, see this table.
For an overview of the preferred formats only (support level 1), see this table.

NOTE: only the file formats that are currently stored are listed in the table. The table will be reviewed and updated on a regular basis.

Specific preservation actions

Each object placed within 4TU.ResearchData will be subject to established preservation techniques in order to maintain its integrity and our ability to reproduce the object as necessary to ensure continuous access over time.

The following actions are undertaken within the preservation process:

Preservation metadata: Each dataset may consist of several files with different file extensions (.pdf, .jpg, .nc etc.).

These extensions are mapped to mime types at ingest. The rationale is that (1) our 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 are indexed in a special field so it is easy to make sets of a specific type that may require a preservation action. Over time, multiple versions of a dataset may be created as formats become obsolete. All version will be kept and their creation date is recorded in the metadata. In addition, there is to the usual bibliographic metadata augmented with relational metadata (many of our datasets have semantic relations with other digital object like other datasets or representations of instruments). This also eases the creation of collections of datasets for a specific preservation purpose.

File format recognition: At the moment, mime types are determined by a fixed 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, and always the possibility to override it manually.

Secure storage and backup: Data storage of 4TU.ResearchData is managed by the IT 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.
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. On regular basis security updates and patches will be installed when approved. These preservation procedures are outsourced to the ICT department of Delft University of Technology and put down in a service level agreement. Development of additional consistency checks are ongoing.

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

Hardware migration: 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 data 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.