DataverseNO Preservation Plan

site.uit.no/dataverseno/about/policy-framework/preservation-policy/preservation-plan/

Purpose and Identification

The DataverseNO Preservation Plan describes the implementation of the DataverseNO Preservation Policy. The preservation plan describes actionable steps to be taken to preserve published Datasets within the DataverseNO repository (the Preservation Action Plan), and documents why and how the Preservation Action Plan was chosen.

The Document History and Version Control Table at the end of this document gives information about the date on which this plan was most recently modified as well as a list of previous approved versions (if any), including the dates defining the periods in which the plans were active.

Status and Triggers

This is the first published version of the DataverseNO Preservation Plan. It has been approved by the Board of DataverseNO and has been active since the date indicated in the Document History and Version Control Table at the end of this document. The creation of this plan was triggered by the wish of the DataverseNO repository to demonstrate compliance with the CoreTrustSeal requirements for trustworthy data repositories.

The following events will trigger a review and potential revision of the DataverseNO Preservation Plan:

- **Changed repository profile:** Changes in the profile of the DataverseNO repository and its collections may require a revision of the current version of the Preservation Plan. Examples for such changes are newly accepted object formats resulting in new types of asset groups (see section *Digital Assets* below), or significant changes in the repository size.
- **Changed environment:** The environment of the DataverseNO Preservation Plan consists of the technical environment, the Designated Community, as well as the owner institution and the partner institutions. Changes in the environment can lead to a change in preferences, for example with respect to the system context in which a preservation action needs to operate. They might also imply a change in factors which influence existing preservation plans, for example changed prices for hardware or software. Other relevant changes are the availability of new preservation strategies or impending obsolescence of object formats which are used in the current Preservation Plan.

- **Changed objective:** Changes and developments in the environment can change the objectives for preservation evaluation over time. In this case, it will be necessary to evaluate the current Preservation Plan against changed objectives. Examples are changes in high-level policies or legal obligations that have an impact on preferences and objectives. Changes in the Designated Community, such as the type of software available to the users or new ways of using the objects of interest, may also affect the goals and objectives.
- **Periodic review:** Periodic reviews of the current Preservation Plan are needed to verify its appropriateness, and to improve and further develop it. Independent of the triggers described above, DataverseNO commits to periodically review its Preservation plan at least every third year.

DataverseNO continually monitors the events described above.

Organizational Setting

The reference frame of the DataverseNO Preservation Plan is given by the DataverseNO Preservation Policy. As stated in the preservation objectives contained in the Preservation Policy, by the requirements of DataCite, DataverseNO commits itself to provide preservation of and access to published research data for a minimum of ten years after the date of publication in the repository. The intention for DataverseNO is, however, to facilitate access to archived data in a long-term perspective, as stated in the Steering document for DataverseNO.

In addition to these preservation objectives, the Preservation Policy also describes a broad range of high-level influence factors and constraints that have an impact on the decisions taken in defining this Preservation Plan.

Digital Assets

The application of the DataverseNO Preservation Plan is based on periodic reviews of the digital objects contained in the repository. The results of these reviews are summarized in the DataverseNO Digital Assets Report. The current version of the DataverseNO Preservation Plan is based on the current version of the DataverseNO Digital Assets Report.

With regard to the DataverseNO preservation program, the current version of the DataverseNO Digital Assets Report divides the digital assets in DataverseNO into the following five Digital Asset Groups:

Name of Digital Asset Group	Brief Description of Asset Group
Group 1	Items with only non-preferred file format(s)
Group 2	Datasets without ReadMe file

Group 3	Container files (.zip or .tar)
Group 4	Files in file formats with unclear preferability status
Group 5	All other assets

Asset Group 1

The research data items in Digital Asset Group 1 are stored only in file formats that are not considered as preferred in the DataverseNO deposit guidelines. These cases of non-compliance are due to the lack of provisions in previous guidelines, single occurrences of insufficient curation, or the fact that the data at the time of initial publication could not be saved in or converted into a preferred file format. More details about this Digital Asset Group are available in the current version of the DataverseNO Digital Assets Report.

Asset Group 2

The research data items in Digital Asset Group 2 lack a ReadMe file. These cases of non-compliance with the DataverseNO deposit guidelines are due to the lack of provisions in previous guidelines or single occurrences of insufficient curation. More details about this Digital Asset Group are available in the current version of the DataverseNO Digital Assets Report.

Asset Group 3

The research data items in Digital Asset Group 3 are stored in container files. The DataverseNO deposit guidelines do not recommend container files. In previous versions of the repository software it was not possible to maintain the folder structure of ingested files. In cases where the folder structure was important DataverseNO has therefore accepted container files preferably of the type .zip or .tar. Since the repository software now supports retention of folder structure, DataverseNO considers to unpack these container files. More details about this Digital Asset Group are available in the current version of the DataverseNO Digital Assets Report.

Asset Group 4

The research data items in Digital Asset Group 4 are stored only in file formats whose preferability status is considered as unclear by the repository management. More details about this Digital Asset Group are available in the current version of the DataverseNO Digital Assets Report.

Asset Group 5

Any assets not contained in any of the Digital Asset Groups 1 to 4 above are part of the residual Digital Asset Group 5. These assets comply with the DataverseNO deposit guidelines. More details about this Digital Asset Group are available in the current version of the DataverseNO Digital Assets Report.

Requirements for Preservation

The preservation of Datasets in DataverseNO follows the regulations and guidelines which UiT The Arctic University of Norway (owner of DataverseNO) applies to at all times. In particular, the DataverseNO Preservation Policy defines the framework of and limits of how Datasets are preserved in the DataverseNO repository.

Preservation Levels

As a main rule, DataverseNO commits to facilitate access to and reuse of archived data in a long-term perspective. However, given common constrains such as cost limitations, different types of digital assets receive different levels of preservation efforts according to the three preservation levels determined in the DataverseNO Preservation Policy. Applied to the asset groups defined in the previous section, the following preservation levels are required:

- Asset Group 1: Preservation Levels 1 + 2
- Asset Group 2: Preservation Levels 1 + 2
- Asset Group 3: Preservation Levels 1 + 2
- Asset Group 4: Preservation Levels 1 + 2
- Asset Group 5: Preservation Levels 1, 2 + 3

Significant Characteristics

According to the DataverseNO Preservation Policy, DataverseNO employs the significant characteristics as defined by Archivematica (link) and summarized for each file category in the overview below. In cases requiring compromise, transformations that maintain the content of the object will be prioritized over those that preserve the presentation and behaviour of the object. In case of uncertainty about the significant characteristics of an object to be preserved, DataverseNO strives by reasonable efforts to obtain advice in the matter from the Depositor or from other representatives of the user group at stake.

File Category	Characteristic	Component
Audio	Duration	content
	number of channels	content

	channel mapping	structure
	sampling frequency	rendering
	bit depth	rendering
Raster Image	image width	content
	image height	content
	sequence of images	structure
	X sampling frequency	appearance
	Y sampling frequency	appearance
	samples per pixel	appearance
	bits per sample	appearance
	extra samples	appearance
Page Layout	textual content	content
	Formatting	appearance
	Layout	appearance
	Bulleting	appearance
	colour and embedded graphics	appearance
	metadata: page count, fonts	
Spreadsheet	all cell data and formula	content
	author, title, etc.	context
	cell locations, the nested worksheets, etc.	structure
	cell formats	appearance
Text	all content	content
	location of every linebreak	structure
Video	imageStreams	video
	audioStreams	audio
	Length	video

	Width	frame
	Height	frame
	bitDepth	pixel
	colourModel	pixel
	colourSpace	pixel
	pixelAspectRatio	pixel
	Framerate	video
	Interlace	frame
	Metadata	

For Web Files, see https://wiki.archivematica.org/Significant_characteristics_of_websites. For the following file categories, DataverseNO does currently not have any overviews of significant characteristics: Container Files, Data Files, Database Files, Developer Files, GIS Files, Settings Files. For Container Files it is most crucial to preserve the significant characteristics of the contained files.

Record Characteristics

The digital assets in DataverseNO must be preserved as both authentic and accessible records, which means that metadata must be kept that support and validate the record characteristics of reliability, integrity, and usability.

To ensure reliability DataverseNO is required to employ transparent and fully documented preservation strategies, and provide metadata required to describe the content, context and provenance of the record.

To ensure integrity DataverseNO is required to provide bit-level preservation and metadata to describe all authorised actions undertaken in the course of content and bit-level preservation.

To ensure usability DataverseNO is required to provide content preservation and the provision of metadata sufficient to allow the record to be located, retrieved and interpreted.

Process Characteristics

The different preservation strategies employed by DataverseNO vary in terms of involved processes as well as the complexity and the degree of effort required for their implementation.

Bit-level preservation (Preservation Level 1) consists of two main processes, bit stream copying and fixity checking. Bit stream copying requires that all information stored in the repository be regularly backed up for use in the event of data loss. Fixity checking requires that all materials stored in the repository be regularly checked for fixity. This means that the MD5 checksum values calculated for all files at a given point in time must be compared with the MD5 checksum values that were created and stored for the corresponding files at their time of ingest.

Normalization (Preservation Level 2) after publication requires that the latest version of published Datasets be inspected in order to detect non-compliance with the DataverseNO Deposit Guidelines. This means that detailed asset list must be created by extracting Dataset and Data File metadata from the repository database including the following information: PID of the Dataset, Dataset version, Dataset title, publication date, date of last update, name of collection where the Dataset is published, name of Depositor, name of curator, file name, file extension, and file category. Datasets not containing any file with file name indicating documentation (preferably indicated by the string “ReadMe” or similar), and files with file extension not listed as preferable file formats in the DataverseNO Deposit Guide must be inspected. Non-compliant Datasets must be amended. This may imply (1) that Data Files in non-preferred file formats are provided in preferred formats, either by format converting or by exporting the content stored in the original software environment to a preferred format, and/or (2) that missing metadata about a Dataset is provided. The files to be normalized must be downloaded and normalized. The normalized versions of tabular files must be fixity checked. In case of deviating Universal Numerical Fingerprint (UNF) checksums, the normalization process must be repeated until successfully completed. In cases of missing documentation, the Depositor must be contacted to request the necessary documentation.

Format migration (Preservation Level 3) requires that the preferred file formats of files in the latest versions of published Datasets be monitored for obsolescence. Files that are in file formats that at the time of publication were considered as preferred for long-term preservation, but later turn out to be in danger of becoming obsolete, must be migrated to new file formats. The new file format may be a new version of the currently employed format, or an entirely new format that is considered to have superseded the currently employed format in terms of suitability for long-term preservation. The files in format(s) to be migrated must be downloaded and migrated. The normalized versions of tabular files must be fixity checked. In case of deviating Universal Numerical Fingerprint (UNF) checksums, the normalization process must be repeated until successfully completed.

Infrastructure Characteristics

The implementation of the preservation strategies utilized in DataverseNO require the employment of infrastructure at the levels of hardware, software, as well as staff and organization.

Being part of the general and automatized routines for IT security and sustainability at the owner institution of DataverseNO (UiT The Arctic University of Norway), the implementation of **bit-level preservation (Preservation Level 1)** is relatively straightforward. DataverseNO is running on UiT's centralized storage and virtualization infrastructure. Any content in DataverseNO is backed up using an enterprise class backup system with retention policies ensuring that multiple copies are maintained of all data in the system. Data recovery is available from backup as necessary. The hardware and software used for backup are required at UiT regardless of DataverseNO, and the backup of DataverseNO does thus not require considerable additional infrastructure. Fixity checking is implemented in a highly automatized and thus scalable way by setting up scripts that regularly calculate the MD5 checksum values of the files stored in the repository and compare these values with the MD5 checksum values that were automatically created and stored at the ingest of the same files.

In most cases, Research Data Service staff operating DataverseNO have access to the IT infrastructure that is required for the implementation of **normalization (Preservation Level 2)** after publication. The creation of asset lists is done by database queries which, once set up, do not require considerable manual efforts to be run and processed. Conversion into or export to preferred file formats requires some manual efforts by DataverseNO Research Data Service staff. The provision of missing ReadMe files may require considerable manual efforts from curators to get in touch with and request these files from the Depositors. It may turn out difficult to provide a missing ReadMe file several years after publication of a Dataset.

The staff and organization efforts required for the implementation of **format migration (Preservation Level 3)** vary according to the file category and file formats used for the different digital objects to be preserved. Spreadsheet Files (.123, .ods, .xls, .xlsx) and Text Files (.docx, .rtf) are expected to be relatively easy to migrate to preferred file formats because Research Data Service staff operating DataverseNO have experience from normalization of objects in these formats. For other file categories and file formats, format migration may require considerable amount of manual effort, especially to evaluate suitable preferred formats. In the case where a large number of objects are considered to be migrated (e.g. .zip and .tar files) the possibility of a scalable migration approach via API should be considered.

At the level of staff and organization, the implementation of the preservation strategies utilized in DataverseNO require – in addition to the operational efforts described above – considerable training efforts, in particular in the establishment phase of the repository. The different Research Data Service staff members responsible for the different parts of the

DataverseNO Preservation Plan must be thoroughly trained in preservation processes and technology in accordance with international and national standards and best practice recommendations.

Choice of Preservation Strategies

According to the DataverseNO Preservation Policy, DataverseNO employs the following Digital Preservation strategies: Normalization, format migration, bit stream copying, and fixity checking. These preservation strategies are applied to digital objects in the repository at three preservation levels according to the type of file format these objects are represented in. Preservation Level 1 covers all objects in the repository, and the applied preservation strategies to be applied are bit stream copying, and fixity checking. Preservation Level 2 covers also all object, and the preservation strategy to be applied is normalization. Preservation Level 3 covers objects in preferred file format(s), and the preservation strategies to be applied is format migration.

Given the clear commitments stated in the DataverseNO Preservation Policy, there is currently no need to further evaluate and decide on what kind of preservation strategies to apply.

Costs

DataverseNO has not attempted to estimate the costs pertaining to the implementation of this first version of the DataverseNO Preservation Plan. Due to the limited amount of data currently published in the repository the repository management have concluded that DataverseNO has sufficient resources to carry out the Preservation Plan. The experiences from the implementation of this first version of the preservation plan will serve as valuable input to the cost estimation of the implementation of future versions of the preservation plan.

Roles and Responsibilities

The roles and responsibilities for carrying out, monitoring, and re-evaluating the DataverseNO Preservation Plan are distributed among the members of the following stakeholder categories as defined in the DataverseNO Preservation Policy: Depositor, curator, collection management, repository management, advisory committee, and board.

The **repository management** has established this preservation plan and takes care of its implementation, monitoring, and re-evaluation, including the following tasks:

- Make sure that bit-level preservation is carried out regularly
- Create asset reports and detailed asset lists for all collections within DataverseNO, and update these reports and lists as necessary

- Create detailed preservation instructions based on the Preservation Action Plan contained in the DataverseNO Preservation Plan
- Inform and communicate with collection managements about how to carry out preservation actions
- If necessary, carry out format migration, in particularly by applying automatized processes
- Assist in and monitor implementation of preservation actions
- Re-evaluate preservation plan
- Establish new version of preservation plan

The **collection managements** are responsible for the implementation of the preservation actions to be applied to the digital assets stored in their collections, including the following tasks:

- Carry out preservation actions as specified by the repository management
- Inform and communicate with the curators of the collection about how to carry out preservation actions
- Communicate with and provide feedback to repository management about implementation of preservation plan

The **curators** of the different collections take care of the implementation of the preservation actions as requested by the collection management, including the following tasks:

- Inform and communicate with Depositors about how to provide necessary information and assistance to carry out preservation actions
- Carry out the processes that constitute the preservation strategies normalization and format migration.
- Communicate with and provide feedback to collection management about the execution of preservation actions

The **Depositors** are responsible for providing necessary information and assistance to the curators to carry out preservation actions.

When presented with (issues regarding) the DataverseNO Preservation Plan, the **advisory committee** for DataverseNO, and the advisory committees for collections within DataverseNO give advice to the repository management about preservation issues.

The **Board** of DataverseNO have the overall responsibility for all aspects of the DataverseNO preservation policy and the DataverseNO Preservation Plan, including the following tasks:

- Give advice to the repository management about the establishment and revision of the preservation plan
- Approve new versions of the preservation plan

Preservation Action Plan

To keep the digital assets stored in the repository usable and accessible in the long term, DataverseNO has defined a preservation action plan containing the following concrete actions to be undertaken by the stakeholders and applying the procedures to the digital assets as defined in the above sections of this preservation plan:

Preservation Action 1

- **Preservation issue:** Data loss.
- **Preservation strategy:** Bit Stream Copying.
- **Preservation action:** Automatized regular backup of all information contained in the repository.
- **Asset Group(s):** All.
- **Time frame:** Daily.

Preservation Action 2

- **Preservation issue:** File damage and/or corruption.
- **Preservation strategy:** Fixity Checking.
- **Preservation action:** Automatized regular comparison of checksum values of all files contained in the repository. In case of deviation(s), revert to a valid backed-up version of the object from a previous point in time.
- **Asset Group(s):** All.
- **Time frame:** Monthly.

Preservation Action 3

- **Preservation issue:** Obsolescence of items in only non-preferred file format(s).
- **Preservation strategy:** Normalization.
- **Preservation action:** If possible, normalize files to preferred file format(s), and add as new version to Dataset.
- **Asset Group(s):** 1.
- **Time frame:** At least yearly.

Preservation Action 4

- **Preservation issue:** Lack of reusability due to missing ReadMe file.
- **Preservation strategy:** Normalization.
- **Preservation action:** If possible, provide ReadMe file, and add as new version to Dataset.
- **Asset Group(s):** 2.
- **Time frame:** At least yearly.

Preservation Action 5

- **Preservation issue:** Container files (.zip or .tar).
- **Preservation strategy:** Normalization.
- **Preservation action:** Consider unpacking container files. If decided to unpack, apply preservation action 3 or 6, and add unpacked files as new version to Dataset.
- **Asset Group(s):** 3.
- **Time frame:** At least yearly.

Preservation Action 6

- **Preservation issue:** File in file formats with unclear preferability status.
- **Preservation strategy:** Normalization.
- **Preservation action:** Review/reconsidering of preferability. In the case of non-preferable formats, apply normalization.
- **Asset Group(s):** 4.
- **Time frame:** At least yearly.

Acknowledgements and References

The references below point to documents and resources which the present document is adapted from and inspired by, or which are otherwise referred to in the present document.

Archivematica Significant characteristics. https://wiki.archivematica.org/Significant_characteristics

Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., & Hofman, H. (2009). Systematic planning for Digital Preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries*, 10(4), 133–157. https://doi.org/10.1007/s00799-009-0057-1

FileInfo file format registry. https://fileinfo.com/

Government of Canada. Digital Preservation Plan Framework for Cultural Heritage Institutions. https://www.canada.ca/en/heritage-information-network/services/digital-preservation/plan-framework-museums.html

Universal Numerical Fingerprint (UNF). http://guides.dataverse.org/en/latest/developers/unf/index.html

York University Audio – Preservation Action Plan. https://digital.library.yorku.ca/preservation-action-plan/audio

York University Digital Preservation Implementation Plan. https://digital.library.yorku.ca/documentation/digital-preservation-implementation-plan

York University Digital Preservation Strategic Plan.
https://digital.library.yorku.ca/documentation/digital-preservation-strategic-plan

York University Environmental Monitoring of Preservation Formats.
https://digital.library.yorku.ca/documentation/environmental-monitoring-preservation-formats

York University Image – Preservation Action Plan.
https://digital.library.yorku.ca/preservation-action-plan/image

York University Registry of file formats.
https://digital.library.yorku.ca/documentation/registry-file-formats

York University Video – Preservation Action Plan.
https://digital.library.yorku.ca/preservation-action-plan/video

Contact researchdata@hjelp.uit.no with questions or to request an addition or revision to this plan.

Plan Document History and Version Control Table

Version	Action	Approved By	Action Date
1.0	Plan issued.	Board of DataverseNO	2019-10-03