Preservation Policy for Edmond

1 Implementing the Preservation Strategy

In order to manage and preserve data according to the requirements, Edmond (https://edmond.mpdl.mpg.de) employs most principles of the OAIS reference model[1] (see below) and related recommendations like FAIR principles. This base allows Edmond to seek compliance with the community's best practices.

Before a data producer starts the depositing workflow, he or she can take advantage of various support offers provided by the MPDL or local research services in the Max Planck institutes. The MPDL offers guides, tutorials and consultancy regarding different aspects of research data management in general and for Edmond in particular[2]. Librarians, research coordinators and scientific IT staff, available in many institutes, can help with data and metadata preparation according to the respective scientific field. Combined with different technical measures (like e.g. metadata validation), these offers remedy the fact that Edmond, as an institutional repository for a huge variety of scientific fields, cannot offer a sophisticated central curation of data. However, Edmond enables the possibility to introduce local curation workflows by applying its flexible role and permission system on dataset level.

During data ingest, Edmond creates a first draft version of a dataset (including a metadata record and one or more data files). It is assigned a draft DOI and is not publicly available yet. This corresponds to the OAIS SIP. The data files are directly uploaded to an S3-compatible Ceph object storage at MPGs computing centre GWDG. Redundant copies are stored automatically on distributed GWDG servers. For every ingested file, an MD5 checksum is automatically calculated and stored together with the file in the object storage as well as with the metadata in an SQL-compatible database. File integrity checks are performed automatically on a regular basis by the Ceph object storage (CRC-cyclic redundancy checks) and can be started manually any time. The metadata database is backed up completely on a daily basis. Additionally, continuous archiving allows point-in-time-recovery of the metadata. These measurements allow the complete restoration of the system including the data. At this stage, metadata and files can still be changed. In addition, it is possible to share the dataset with other users, which allows curation or review workflows.

Once the dataset is ready and complete, it can be published. A technical validation ensures that all required metadata fields exist. Integrity checks via checksums are performed again for files under a certain size, all others are checked with CRC on a regular basis (see above). The metadata can automatically be exported to different community standards like Dublin Core, DataCite and JSON-LD (schema.org). The reserved DOI is published together with a latest version of the DataCite metadata export, which makes it globally link- and searchable. Every change on metadata and/or files would now result in a new version of the dataset, the original version is kept. Thus, any alterations are accurately documented and ensure the authenticity of a dataset. Deaccessioning (public access removal) is still possible for important reasons and can be done by the Edmond support team.

Edmond does not create separate AIPs, which sets it apart from OAIS. Instead, the dataset is prepared as DIP at this stage. The metadata exports are stored in the object storage next to the files in text format. Datasets are distinguished by a prefix structure, which is equal to the DOI. This allows human-

[1] https://public.ccsds.org/pubs/650x0m2.pdf.
[2] https://edmond.mpdl.mpg.de/guides/help.html and https://rdm.mpdl.mpg.de/.

and machine-readable dataset “packages”, which can be accessed using the widespread S3 protocol standard independent of the repository software (Dataverse).

Datasets can only be published when they provide certain required metadata fields for citation and findability. Edmond sitemaps and structured data (JSON-LD) for each dataset can be used by third party catalogues and search engines and result in a high dissemination.

2 Monitoring Processes

	Description	Process Duration	Responsibility
1	Monitoring of the technical infrastructure and securing error-free operation	Continuously	MPDL System Administration & IT-Support, GWDG System Administration
2	Monitoring of potential external threats for the IT infrastructure. Monitoring and installation of security updates. Frequent penetration tests of the services.	Continuously. Additionally, yearly update of the security guidelines by the IT Security Teams of the MPDL and MPG	IT Security Teams of MPDL and MPG. System Administration of MPDL & GWDG
3	Monitoring of the Dataverse Software Community for new developments and software updates.	On a regular basis	MPDL Development Team
4	Monitoring of the repository‘s community by providing workshops and events. Communication and networking to get direct feedback from the community regarding changes on the repository‘s requirements.	Continuously	MPDL Edmond Service Manager & Support Team
5	Monitoring of the MPG‘s internal and external regulatory and legal requirements. Integration into the project documentation.	Continuously	MPDL Edmond Service Manager & Support Team

3 Contact

You can always contact the Edmond support team via edmond@mpdl.mpg.de.

Please also see the available Guidelines for Long-term Archiving for Edmond (https://hdl.handle.net/21.11116/0000-000C-A18B-1).

Last edit by 22 February 2023.