European Funds
Digital Poland

Republic of Poland

European Union
European Regional Development Fund

Preservation Policy

Preservation Policy

This document describes the commitments and approach of the MOST Wiedzy Research Data Catalog to manage published datasets in the long term responsibly.

Aims and scope

Since 2016, Gdańsk University of Technology has implemented a project called Multidisciplinary Open System Transferring Knowledge. The result of this project is a platform of the same name which aims to provide free access to the resources created and gathered at the University.

Creating the MOST Wiedzy Open Research Data Catalog ensures continuity of the research data life cycle for the scientific output produced and collected by the designated research community. A suitable infrastructure that allows adding metadata description and longtime data storing is crucial to make different stages of this cycle, such as data preservation, dissemination, or reuse of data effectively.

Objectives of the Repository concerning this policy:

1. ensuring long-term access to research data
2. maintaining continuous stability of the repository operation
3. ensuring the authenticity, integrity and security of published data.

Organisational and financial maintenance

Institutional commitment

The MOST Wiedzy Research Data Catalog is located at the Gdańsk University of Technology, which holds all copyrights to this project. A university is a public unit - subordinate to the Ministry of Education and Science. It can be assumed that it is a financially stable institution that is not threatened with closure or liquidation.

As a project leader, the Gdańsk University of Technology is responsible for repository infrastructure development, data preservation, and metadata quality. It provides the know-how needed to share research data in line with FAIR principles.

The specific roles in the Repository are performed by:

- IT Services Center at Gdańsk University of Technology - managing the Repository from the technical point of view and developing new features;

- CI TASK - providing secured storage;
- Gdańsk Tech Library (Open Science Competence Center) – providing support for the users (expertise role) and contributing to the development of the repository features.

Financing

The Repository receives long-term financing from public funds and Gdańsk Tech's funds. It, therefore, offers the stable long-term operation of the entire infrastructure and support of employees responsible for the Repository.

Multidisciplinary Open Knowledge Transfer System - stage II: Open Research Data' co-financed by the European Regional Development Fund under the Operational Program Digital Poland 2014-2020. The data repository was created as part of the project 'Bridge of Data. The received financing requires a durability period of 5 years, starting from the end of the project. This means that the project outcomes - including the data repository- must be preserved until the end of 2026. The Gdańsk University of Technology undertook the responsibility to fulfil this obligation.

The MOST Wiedzy Research Data Catalog's role in implementing the long-term scientific development strategy of the Gdansk University of Technology, the rector of Gdansk Tech, issued a statement declaring taking over the responsibility for the Repository.

Preservation Strategies

The Repository meets the requirements to be completed to ensure the long-term storage of published data:

- the data published in the Repository shall be accompanied by appropriate documentation enabling their use and reuse;
- data is verified, validated and supervised according to predefined workflows.
- the data is described and enriched with metadata following standards and best practices;
- datasets and metadata are stored for a period of time not shorter than ten years;
- the authenticity, integrity and reliability of the data sets retained for future use shall be preserved.

The Repository maintains its continued commitment to the FAIR Principles to make the data findable, accessible, interoperable and reusable, as described in our Mission Statement.

Data verification

The Open Science Competency Center verifies the entire dataset and decides whether it is ready for publication in the Repository or needs additional improvements from the author. If necessary, the dataset is returned to the author for correction.

Open Science Competence Center supports researchers on every stage of the data life cycle and over-preserved data regarding its compliance with FAIR principles. These activities include correcting metadata, improving and standardising descriptions, helping in data versioning, and converting data to new formats for reusability improvement.

Data storage

The entire infrastructure of the repository works (hardware, software, storage media) is maintained under applicable best practices. Infrastructure and services are updated regularly. Data storage systems are constantly renewed, minimising media degradation risk.

The entire repository infrastructure is distributed over two private computing clouds. The frontend and business logic layers are located in the IT Services Center server room at the Gdańsk University of Technology. At the same time, the datasets (files) themselves are stored on CI TASK servers. Server rooms are protected against power failures, fire outbreaks, and air condition system breakdown. They have redundant power lines and own power generators (UPS).

All data is archived using a backup system in multiple copies. CUI and TASK have developed procedures for restoring data from backups if necessary.

Formats migration

Repository recommends open file formats usage. Still, whether the conversion from the original format could result in data or metadata loss, the author of the data should decide on the deposited file formats.

Repository managers, implementing the prevention of the obsolesce of data formats policy, declare an audit of the file formats used in the datasets. The audit will be performed at least once a year. If out-of-date formats are recognised, a decision may be made to convert the selected formats to the current versions. The conversion will be performed programmatically by the Repository's IT administrator team or - in justified cases - by the data stewards team. The result will be assigned a new DOI and placed in the Repository using the existing versioning mechanism. The README.txt file will be attached to the latest version, describing the conversion method and its parameters. The new version will point to the previous version. The audits will omit datasets with assigned licenses, which decline the community of any data modifications (like CC BY-NC-ND license).

Maintaining availability

The Open Science Competence Center verifies deposited data before publication in repository data regarding its compliance with FAIR principles and supports researchers at every stage of the data life. These activities include correcting metadata, improving and standardising descriptions, helping in data versioning, and converting data to new formats for reusability improvement. In the end, the researcher has the final decision on the elements the dataset consists of, its size and formats. 'MOST Wiedzy – Open Research Data Catalog Policy and Information' includes a recommendation for open file formats, which should be used unless there is a risk of information/metadata loss during the conversion process [1]. Public documents were created to raise the awareness of depositors, like instructions for uploaders [2] and the Open Science Competence Center webpage [3].

Thanks to private computing clouds, all system elements have multiple instances. That enables traffic balance, and if one node in the cluster fails, the other nodes continue to operate as usual, providing services to users. The Repository's components are deployed on the server as virtual containers based on the Docker solution. These images are managed by the Kubernetes platform (Kubernetes). Kubernetes has mechanisms that detect node failures and restart to restore the normal operation mode. This combination of technologies allows the provision of high-level service availability.

To increase service availability, the administrators monitor these private computing clouds by using applicable external services capable of detecting the repository unavailability and notifying the IT Team by sending e-mail / SMS messages of this fact.

Data validation

All datasets in the Repository are subject to regular audit verification, i.e. comparisons of checksum values calculated at a given point in time with those generated at the datasets' time of ingest

The Repository presents the calculated S3 ETag for each dataset and the verification algorithm. S3 Etag is calculated after the uploading process is completed. This allows the data depositors to verify whether the Repository's datasets version corresponds to the local copy. Anyone downloading datasets from the Repository can perform a similar action to validate data integrity.

Checksums are also used to monitor whether files have been corrupted. Such a mechanism allows the identification of damaged or lost content and restoring the correct version from backups.

Security

The MOST Wiedzy Research Data Catalog has multi-level access security. It is integrated with security levels of mostwiedzy.pl, which uses a proprietary authorisation system. The user authentication is based on external ID providers (such as PIONIER.Id, ORCID or Gdansk Tech Central Authentication Service) using OUATH 2.0 protocol.

Infrastructure inspections are carried out regularly to ensure a high level of security and stability of services. An external auditing company also carried out an audit of the platform's security. The audit confirmed the correct implementation of the provided solution. Security audits will be carried out regularly in the future.

Control procedures/version changes

The Repository declares that the published data is unchangeable. After dataset acceptance from the Open Science Competence Center, it is closed from being edited/deleted. Only designated people with strict access rights (repository administrators) can delete the data. The authors of the dataset can provide changed/new files only by creating a new version of the dataset. Creating a version generates a new and unique DOI assigned to the latest version. This mechanism keeps all versions from history with their DOI numbers. The "old" DOI number remains active all the time.