Long-term preservation policy
of the University of Warsaw Research Data Repository

This is the English translation of the Long-term preservation policy of the University of Warsaw Research Data Repository. The original text in Polish can be found here.

1. Organisation of the Repository

1.1 Responsible entity

The University of Warsaw Research Data Repository ('Repository') is operated by the University of Warsaw, with its registered office in Warsaw (00-927) at Krakowskie Przedmieście 26/28 ('UW') via the Repository website (https://danebadawcze.uw.edu.pl/).

1.2 Purpose of the Repository

The purpose of the Repository is to make research data, such as tabular data, images, audiovisual materials, and any other type of data produced, collected, or described for the purposes of conducting scientific research at the University of Warsaw or in cooperation with the University, available on the web.

1.3 Scope of data and target audience

The Repository is an institutional repository of research data at the University of Warsaw, maintained for all members of the UW scientific community.

Research data from all fields of science may be collected, stored and made available in the Repository.

1.4 Basis for the Repository's operation

The Repository operates according to the Terms of Use available at https://danebadawcze.uw.edu.pl/terms-of-use-page.xhtml. Additional information on depositing research data is provided in the "User Guide" available at https://repod.icm.edu.pl/guides/en/1.0.0-SNAPSHOT/user/index.html and on the Repository's information website available at http://info.danebadawcze.uw.edu.pl.

The Repository was created as part of the "Excellence Initiative–Research University" program, implemented by the University of Warsaw between 2020 and 2026.

1.5 Repository staff

Tasks related to the day-to-day operation of the Repository and research data management in the Repository are set out in documents defining the responsibilities of relevant employees of the Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw (ICM UW).

Responsibilities related to the day-to-day maintenance of the Repository are taken into account by the ICM UW management when defining the organisational structure of ICM UW (in accordance with §5 of the ICM UW Regulations) and in the workforce planning process.

1.6 Financing of the Repository's activities

The development and maintenance of the Repository until 2025 was financed from the "Excellence Initiative–Research University" program funds.

The University of Warsaw Open Access Policy and the University of Warsaw Strategy 2023–2032 identify the Repository as a component of the University's research infrastructure, to which convenient, integrated and standardised access must be provided, and highlight the direct relationship between the implementation of this commitment and the need to take into account the increasing cost and complexity of research data management.

1.7 Information on data usage
The system collects information on downloads of individual files and datasets. Information on the number of downloads of individual files is publicly available.

Information about the total number of downloads of files and entire datasets is also publicly available for the entire Repository collection.

1.8 Audit and certification
In June 2025, the process of applying for the Core Trust Seal certificate for the Repository began.

1.9 Proceedings in the event of termination of the Repository
In the event of circumstances necessitating the termination of the Repository, the University of Warsaw will seek to transfer all the data stored in the Repository to another location, while maintaining the continuity of the correct functioning of the DOI numbers assigned to the datasets.

The choice of a new location for the datasets will depend on the current availability of infrastructure enabling data transfer.

1.10 Amendments to the long-term preservation policy
Subsequent versions of this document can be found on the Repository's information website at http://info.danebadawcze.uw.edu.pl.

2. Repository operations and data management
2.1 Collections
Within the repository there are collections of all UW faculties and other units, such as centres and institutes. The principles of operation of the collections are the same as those of the main collection.

2.2 Data acquisition
The Repository does not define specific requirements for the properties of datasets, except for the requirement to deposit at least one file within a dataset. Open file formats are preferred.

The Repository always retains the original format of the deposited file..

The Repository accepts datasets of any size, taking into account the technical accessibility of datasets of a specific size for users. The size limit for a single file is 5 GB.

An MD5 checksum is calculated for each file uploaded to the Repository. This makes it possible to compare it with the checksum of the file calculated on the User's side and verify the correctness of the data transfer.

The Repository allows data files to be uploaded through a web-based graphical user interface or API.

When creating a dataset, the User initially produces a draft version, which may be subject to changes and additions in terms of metadata and files. Once it has been published and assigned a version number, the User cannot modify the metadata and files that comprise it. However, the User can create additional numbered versions of the dataset based on the existing, published version.

2.3 Requirements for depositing datasets
The Repository does not charge fees to users depositing or downloading data.

The use of the data storage and sharing functionality of the Repository is possible for persons with an active account in the UW Central Authentication System (CAS UW) or, in the absence of such an account, with an account to which an e-mail address and authentication password are assigned, with the appropriate permissions granted by the Administrators.

The scope of information required to create and edit a dataset is indicated directly in the metadata form. A dataset can be submitted for verification if it contains at least one file and all required metadata fields have been completed.

The configuration of metadata sets and mandatory fields is done at the Repository level or a collection level. The configuration is carried out by the Repository Administrators.

2.4 Minimum set of metadata in the Repository
The Repository, as a general-purpose repository, requires only basic information about the datasets to be entered. These metadata fields are:

- title;
- author (last name and first name);
- contact person (e-mail);
- description (content);
- keywords (content);
- subject area.

In addition, the following fields are available in all collections of the Repository and are optional:

- title translation (content; translation language);
- author (identification system; identifier; affiliation; ROR of affiliating institution);
- contact person (last name and first name; affiliation);
- description (date);
- MEiN 2022 discipline;
- UNESCO, OECD R&D field;
- keywords (vocabulary; vocabulary URL);
- related publication (citation; relation type; identifier type; identifier; URL);
- notes;
- language;
- metadata language;
- data language;
- documentation language;
- grant information (funding institution; abbreviated name of institution; ROR of funding institution; grant program; grant number);
- related dataset (citation; relation type; identifier type; identifier; URL).

More optional metadata fields are available to be filled in edit mode once the draft version of the dataset has been saved.

Separate requirements for mandatory metadata fields may apply within individual collections.

When depositing data, the User must also indicate the licences or conditions under which individual files will be available.

2.5 Metadata validation
Entering the minimum required metadata is necessary to create a draft dataset and then publish it. The repository software automatically verifies the completion of the relevant fields when the draft version is saved. In addition, the correctness of the description is checked when the dataset is verified before publication.

2.6 Embargoed datasets
For embargoed datasets, the User specifies the date from which the files in the dataset become available. From the moment such a dataset is published, its metadata is publicly available. An embargo can only be set for datasets that do not yet have a published version.

2.7 Maximum embargo period
The maximum embargo period in the Repository is 36 months.

2.8 Files made available with access restrictions
In the case of files in datasets made available with access restrictions, the Repository allows requesting access to a specific file. The request is addressed to the Repository User who deposited the dataset containing the file. The Repository staff receives a copy of the email containing the request for access to this type of resource, but does not interfere in further correspondence between the user wishing to access the files in the collection and the user who deposited it.

2.9 Anti-malware check
When uploading to the Repository, files are subject to anti-virus checks.

If an anomaly is detected, saving the file is prevented, and the User is informed that the file upload operation has failed.

2.10 Verification of deposited datasets
Once research data has been deposited and described in the Repository, the User submits it for verification. Verification is performed by the Administrator. It consists of:

- verification of the correctness of the metadata entered;
- validation of files.

This process (from the moment of submission for verification until the decision to publish or return the dataset for correction) usually takes up to 3 working days.

Formal confirmation of the publication of the dataset version is sent to the depositor's e-mail address.

2.11 Dataset versions
Each dataset deposited by a User consists of at least one version. A dataset version consists of metadata and files containing the data and its supporting documentation.

When creating or editing a dataset, the user initially creates its draft version. This version can be edited in terms of both metadata and files. The possibility of editing is blocked when a specific dataset version is published. An exception to this rule are minor corrections of obvious errors in the metadata applied by the Administrator. Such changes do not require the publication of a new dataset version.

When publishing a dataset, the Administrator determines whether the changes require publishing a draft version as a new major or minor version.

A draft version in which the files have been modified can only be published as a new major version.

The dataset version number is part of the suggested citation of the dataset, which is visible on the dataset page.

2.12 Deletion and withdrawal of datasets and versions of datasets
Once a version of a dataset has been published, it cannot be deleted. However, it is possible to withdraw a version of a dataset in exceptional cases. Withdrawing an entire dataset involves withdrawing all of its versions.

If a version of a dataset is withdrawn, only basic information about the dataset (known as a tombstone) remains publicly available:

- citation;
- reason for withdrawal.

The full metadata and data of the withdrawn version remain available for viewing to Users with system roles permitting to publish and withdraw the resources (Administrators).

2.13 Form of data archiving
Data shall be archived and made available in the form provided to the Repository by the User. The Administrator may request corrections or additions to a dataset before publishing it and, in the case of minor and obvious errors (e.g. typos), make the necessary corrections themselves.

In the case of selected tabular data formats, copies of these data are additionally created in other formats in order to increase the availability of data to Users using different types of software. This conversion is performed automatically. The deposited file is always retained in its original format.

In the case of tabular files, a UNF (universal numerical fingerprint) value is additionally generated, allowing Users to verify the correctness of the conversion.

2.14 DOI identifier
Each dataset deposited in the Repository is assigned a DOI identifier within the Repository prefix 10.58132. This identifier is reserved within the Datacite ecosystem infrastructure when the first draft version of the dataset is created. The DOI identifier is activated when the first version of the dataset is published.

The DOI identifier of a dataset is part of the suggested citation, which is visible on the dataset page.

2.15 DOI identifier and dataset versions
All dataset versions have the same DOI identifier, with the version number serving as their distinguishing feature. Information about all versions of the data sets is presented on the data set page in the “Versions” tab.

Within the metadata, it is also possible to identify other objects (publications, datasets) related to the dataset and the DOI identifiers of these objects, as well as to specify the type of relationship between them.

2.16 System logs

Every file download is logged for all Repository resources.

For each activity of this type, the logs record information about:

- file identifier;
- type of download;
- e-mail of the User downloading the file;
- the name of the User downloading the file (if applicable);
- the position of the User downloading the file (if applicable);
- the date and time of the file download;
- the ID of the logged-in User (if applicable);
- data file identifier;
- dataset identifier;
- dataset version identifier;
- guestbook identifier (if applicable).

2.17 User support in the Repository
The Repository allows Users to contact it via e-mail (danebadawczeuw@icm.edu.pl) and the contact form on its website. Reported issues are resolved by the relevant ICM UW staff.

For questions about a particular dataset, the Repository allows Users to contact the persons indicated as contact persons for matters relating to that specific dataset.

In addition, Users with appropriate system roles receive automatic e-mail notifications regarding:

- dataset creation;
- dataset submission for verification;
- dataset return for correction;
- publication of a dataset;
- submission of a request for access to a restricted file;
- assignment of a system role;
- withdrawal of a system role.

3. Legal aspects
3.1 Rights over datasets
It is the responsibility of the User of the Repository to clarify the legal status of the deposited data under the Terms of Use. By submitting data to the Repository, the User declares that they have sufficient rights to deposit data and make it available. Clarifying any doubts about the legal status of the deposited data under the Terms of Use shall be the responsibility of the User providing the data to the Repository.

3.2 Licence for the University of Warsaw
By submitting data to the Repository, the User grants the University of Warsaw a non-exclusive licence covering the preservation, storage, and reproduction of research data by digital technology for the purpose of operating the Repository. This licence covers all submitted research data, especially metadata, rights related to research data as a collection of files, and files constituting research data. The licence also covers sharing research data with third parties, except where access to a particular file has been limited.

3.3 Declaration regarding metadata and collection of files
By submitting data to the Repository, the User subjects the data to the terms of Creative Commons 0 deed (“CC0”), with its full text available at https://creativecommons.org/publicdomain/zero/

1.0/legalcode, the subject of which is metadata describing the research data and any rights related to the research data as a collection of files.

The inclusion of metadata describing research data in this declaration enables the free exchange of metadata with other services aggregating information about available datasets.

The inclusion of rights related to the dataset as a set of files in this declaration (e.g. collective work rights, sui generis database rights) facilitates the assessment of the extent of the User's rights, which in this situation result only from the rules defined for the individual files included in the set (e.g. the indicated CC licence).

3.4 Conditions for file sharing
For each file included in a deposited dataset, the User can assign a licence from the list of licences available in the Repository, make it available on a fair use or restricted use basis.

In the case of restricted-use files, the decision to make a file available to a specific User is made on a case-by-case basis by the depositor and may be subject to the User's acceptance of additional conditions (e.g. that the file may only be used for research purposes).

3.5 Multiple licensing
It is permissible for the same files to be available under different licences in different versions of a dataset (multiple licensing). If each of these files is publicly available, the User of the file can choose which of these licences to adhere to.

3.6 Form of licence
When the User makes resources, including files and metadata, available under a non-exclusive licence, its written form is not required (Polish law allows for the granting of non-exclusive licences in any form).

When the User makes the dataset files available based on fair use, sharing and using the data is possible under the applicable law. Therefore, in this case, there are no provisions for which it would be necessary to consider the issue of their form.

3.7 Restricted file sharing
If a User makes dataset files available in a restricted manner, the additional conditions for access and use of the files are described for each file made available under these rules.

The granting of access to files to a particular User is in each case a decision of the depositor, who is also responsible for any additional measures to verify the person who requested the data.

3.8 Communication on the use of data
The Terms of Use of the Repository contain information on the terms and conditions under which a User can make data and metadata available. Each User accepts the Terms of Use when registering an account.

Information on the licence under which each file in the dataset has been made available is presented next to each file in the "Files" tab of the dataset page and additionally on the page of each file.

3.9 Monitoring of data use
The Repository does not monitor the compliance of the data's use with the licences under which it is made available.

The Repository does not monitor whether the data downloaded is used in accordance with its licence. In the case of a breach of the licences or access conditions, it is up to the rights holder of the affected resource to take possible legal action.

The Repository also does not interfere with additional User verification for files made available under restricted access conditions (e.g., for research purposes only). This possibility is also left to the Repository Users depositing restricted access files.

No login or account is required to access datasets made available under other conditions.

4. Data storage

4.1 Data integrity and backup copies

The software of the S3 storage on which the data files are stored is responsible for maintaining their integrity and correctness.

In addition, in the event of a failure or accidental deletion of resources from the S3 storage (e.g. as a result of a software error), it is also possible to restore lost resources from an additional backup copy stored in a separate location.

Database and file backups are stored separately.

4.2 Verification of data integrity

When a file is uploaded to the Repository, its checksum is generated. The User can compare it with the checksum generated locally to confirm that the copy of the file uploaded to the Repository matches the file on the User's side.

The Repository periodically checks that the resources that, according to the information in the Repository database, should be in the S3 storage are there indeed. If a discrepancy is found, a backup copy of the files residing in a separate storage in a different location is retrieved.

Periodically, the checksums in the Repository database are also compared with those stored in the files' metadata in the S3 storage.

Checksum information is also displayed next to each file on the dataset and file pages. This makes it possible to compare the match between the downloaded copy of the file on the User side and the copy deposited in the Repository.

4.3 Relation between archival (AIP) and dissemination (DIP) copies

The Repository does not create archival copies (AIPs) separately from copies of datasets subject to dissemination (DIPs).

5. Technical infrastructure

5.1 Repository software

The Repository runs on the free Dataverse software version 4.11, modified by the ICM UW development team. The modified version of the software is called Fairchive, is open source, and its source code is available on the GitHub platform at https://github.com/CeON/fairchive. This platform allows the monitoring and reporting of changes and corrections made to the software code.

5.2 Data storage

File data is stored in an S3 storage. Dataset and file metadata are stored as a relational database in a storage resource attached directly to the server.

The security of the data deposited in the Repository is further ensured by the native redundant storage mechanism of the S3 disk storage on which the data is placed. The system is supported by an external company.

In addition, the metadata database and file data are transferred to a storage resource in another geographical location. This makes it possible to recover data files and the Repository database in case of damage or failure.

File data and the metadata database are backed up once a day. Copies of the database from the last 30 days and copies of data from the last two days are stored in a separate location.

5.3 Proceedings in the event of data disintegration
In the event of a difference in the checksums of the deposited files stored in the database and the system metadata in the S3 storage, the system sends a notification to the Administrators. The Administrators take action to restore system integrity.

5.4 Operation logs
System logs are created and stored in the Repository for the following operations:

- creating a User account;
- withdrawing a dataset version;
- setting an embargo date for a dataset;
- activating a lock on a dataset;
- assigning a system role;
- creating a collection;
- creating a guestbook entry;
- creating a new dataset;
- creating a private URL for a dataset;
- creating a dataset template;
- deleting a draft version of a dataset;
- deleting a draft version, which is the only version of a dataset;
- deleting a dataset template;
- publishing a dataset;
- fetching JSON ID;
- fetching a private URL of a dataset;
- saving a provenance file;
- publishing a dataset;
- publishing a collection;
- removing a lock on a dataset;
- issuing a request for restricted access to a file;
- returning a dataset for correction;
- revoking the system role;
- submitting a dataset for verification;
- editing a dataset thumbnail;
- editing a draft dataset;
- creating a draft version of a dataset;
- editing a collection;
- changing a default role for a User creating a dataset in a collection;
- editing a dataset template;
- editing collection appearance;
- logging a User in;
- logging a User out;

- issuing a request to change the password;
- sending a request to change the password;
- retrieving available login methods from the database;
- changing the maximum length of an embargo;
- activating or deactivating file sharing conditions (restricted access and permitted use);
- changing User account parameters.

5.5 Periodic review of functionalities

Once a year, the Repository's functionalities are reviewed against identified new User needs, technologies, and features. New functionalities will be progressively developed and implemented as staff and financial resources allow.

Document version: 1.0.0