Data Collection Policy

1. Mission

4TU.ResearchData is a data repository for publishing technical-scientific research data and software.

Once published, data are stored in a permanent and sustainable manner, according to the guidelines of the international CoreTrustSeal. As a trustworthy data repository, we demonstrate to researchers that we take appropriate measures to ensure the long-term availability and quality of data we hold.

Our mission is to enable researchers from science, engineering and design disciplines to make a global impact with research data and software by ensuring the accessibility of their published data and software.

The Repository is operated by the TU Delft Library on behalf of the 4TU.ResearchData Consortium. The consortium is open to organizations that share our mission and are willing to make an active contribution to it.

2. Scope

4TU.ResearchData is primarily focussing on data and software from all fields and subjects in science, engineering and design, but also covers subjects in the life sciences (see the appendix for a nominal list of subjects).

Atmospheric and environmental research is one of the areas of interest of 4TU.ResearchData. In fact, data from these disciplines currently form the bulk of the contents in our archive. Most of these datasets are coded in netCDF, which is both a data model and a data format which is very efficient for multi-dimensional array-oriented data. The format is self-describing, i.e. it includes general metadata as well as detailed metadata about variables, dimensions and units used, in a fully machine-readable way (as opposed to, say, a spreadsheet with column headings, which is not really machine readable).

Access to netCDF data (and HDF5) is further enhanced by serving the data via the OPeNDAP protocol. A major advantage of using OPeNDAP is the ability to retrieve subsets of files without the need to download the whole dataset, and also the ability to aggregate series of data files, e.g. a time series, into one 'virtual' dataset.

In terms of the size of the datasets, our focus is on sharing the richly diverse and heterogeneous small datasets, so-called long-tail data. Projects that produce large amounts of data ('Big Data') are the exception and often have their own subject-specific data services in place. Our focus is not to store large datasets comprising petabytes of

data, but to manage multiple data objects in a way that facilitates their reuse.

A dataset is defined here as a group of data files, usually numeric or encoded, related to a specific topic and collected for a specific purpose. It may include both data and the means to generate, interpret or validate data, such as computer models and software code. A dataset consists also of documentation files, such as codebooks, user manuals, workflows, protocols, methodologies, etc., that supports its use or analysis.

Datasets will be published if one of the following conditions apply:

- The dataset has long-term value, for example, to support research re-use, teaching, decision-making or policy formulation, and the dataset can be made available for others to use.
- The dataset supports or will support publications, and the dataset can be made available for others to use.
- There is a funder/institutional policy, legal or contractual requirement to preserve and share the dataset.
- The dataset is finalised for making public, and provided with complete metadata. See the Deposit guidelines for more details.

Where materials fall outside of scope, efforts are made to determine whether another archive or data repository is more appropriate for dissemination or preservation of these data.

There are also criteria for not accepting data:

- *Legal and ethical issues*: where there are insurmountable rights management issues e.g. consent, IPR, copyright and data protection issues which are unable to be satisfactorily resolved and where full use of the data would not be possible without infringing legislation.
- *Lack of sufficient contextual materials to enable re-use.*

3. Data formats accepted

4TU.ResearchData prefers data in a readily usable format, accessible in a variety of computing and technological settings.

4TU.ResearchData prefers data formats that promote easy access and use without compromising research value.

Data in obsolete, proprietary, or hard-to-use formats may still be accepted by 4TU.ResearchData, although these characteristics may compromise any future use of the data other than as-is, bit-level access.

You can read more about our preservation strategy in our Preservation Policy.

4. Removal of datasets

Once it has been published, 4TU.ResearchData will not remove a dataset with a DOI, as it may have been cited by other researchers.
However, if serious grounds exist (for example in case of data falsification or even fabrication) 4TU.ResearchData can withdraw the dataset from the repository, or restrict or prevent access to the dataset on a temporary or permanent basis. Attempts will be made to inform the depositor in such cases.

In the unlikely event of a dataset being removed from public view, a landing page is still accessible from the DOI with the reason why the dataset is no longer available.

5. Licensing

When data is published, a licence will be attached to it to inform users how they may use the data. When depositing data or software for publication in 4TU.ResearchData depositors are required to select a licence from a predefined list.

4TU.ResearchData offers the full range of Creative Commons licences for data, and for software and code, the most frequently used open source licences are supported.

4TU.ResearchData has adopted CC0 (Creative Commons Zero) as the default means for researchers to share their datasets to make its reuse as easy as possible without any legal barrier. If there are reasons or circumstances where data can't be shared with a CC0 licence, depositors can choose another, more appropriate licence for their data.

Guidance on all licence types offered can be found here.

6. Versioning

A new version of a dataset should be created when an existing dataset is reprocessed, corrected or appended with additional data. Modifying the dataset title, authors, licence or the file(s) associated with the dataset will automatically create a new version of the dataset and its DOI.
Updating any other metadata field (description, categories, keywords, etc.) will not generate a new version.

The original DOI of the published dataset will always take you to the latest version of the item. Versions are listed and accessible in the drop down menu under the dataset title.

7. Confidentiality and privacy

4TU.ResearchData prefers data that can reside in the public domain.
Data that is personal, confidential or sensitive in nature, is only accepted for archiving when the data has been anonymised so that individuals cannot be identified.
Pseudonymised data (with the exception of sensitive personal data and special categories of personal data which cannot be deposited in 4TU.ResearchData) can be shared under restricted access conditions.

The matrix shows the different type of data and risk classifications, with examples of data that fit into each classification (not exhaustive):

Type	Characteristics	Examples	Access level	Risk level
Public data	Data that can be freely used, reused and redistributed by anyone with no restrictions on access or usage	Non-confidential information. Anonymous or de-identified information. Identifiable information that a human subject has consented to make publicly available.	Open access	LOW
Confidential data	Data is not generally available to the public	Contractually protected data. Research that has not been completed or finished. Data with commercial potential. Data resulting from sponsored or collaborative research. Pseudonymised data (with the exception of sensitive personal data and special categories of personal data)	Embargoed access / Restricted access; An embargo period can be applied to provide delayed access to the data, or the access level 'Restricted access' can be applied by which only authorized users are granted access to the data.	MEDIUM
Sensitive personal data, special categories of personal data, data pertaining to criminal matters or national security data	Information about a person's religion, race, health, sexual life, political preference, as well as genetic and biometric data, data relating to criminal convictions and criminal offenses or related security measures.	Medical data. Individually identifiable financial or medical information. Biometric data. Certain types of nuclear research.	Closed access – not supported; Instead a metadata only record can be created, but the actual files need to be safely and securely stored at a different location.	HIGH

8. Reuse

Since 2010 the data repository of 4TU.ResearchData has been managed as a resource for researchers in science and technology to deposit and share their data, and for other researchers to download and use data in their research.

Starting with a data collection of hydrological measurements resulting from the DareLux project, 4TU.ResearchData has grown to become what we believe is the largest data repository of its type in The Netherlands.

We encourage researchers to contribute to the data repository by helping them to create metadata for their datasets and by providing an easy to use online upload service.

Anyone using data from the 4TU.ResearchData repository, is expected to cite or reference this work as they would any other scientific research. Even if the licence does not explicitly require users to do so. Citing data is considered good scientific practice and helps to avoid charges of plagiarism.

In order to cite 4TU.ResearchData datasets properly, we provide a ready-to-go citation for each dataset that authors can use.

4TU.ResearchData measures the uploads, downloads of the data, page views, citations, registered users, and unique logins.

Another way we believe sharing and reuse of datasets can be encouraged, is by providing adequate data services in the data collection and data processing/analysis phase of the research process.
For this purpose we are exploring research collaboration platforms that can be used for storing, processing and sharing dynamic research data, tools for data visualization, and the use of electronic lab notebooks for recording research.

Consistent documentation of research methods, calculations, and results during the research, will in the end help publish or otherwise share research when others want to reproduce and reuse what has been done. As data sharing and reuse are the main goals of our data repository, we will do all we can, and are eager to get involved, in establishing services that support these goals.

9. Charges

Every researcher, both in the Netherlands and abroad, can upload up to 5 GB of data per year to our data repository free of charge. For depositing additional data, there is a one-off cost of € 4.50 per GB.
For researchers from member institutions of 4TU.ResearchData, different costs and upload limits apply. Please check the most up-to-date information in the Costs tab on this page.

There might be circumstances in which the extra costs can be reduced or waived. 4TU.ResearchData staff are happy to discuss different possibilities for larger collections of research data.

By default, published datasets can be downloaded free of charge and be stored for a minimum of 15 years.

10. Responsibilities

4TU.ResearchData staff will determine whether datasets submitted to the repository for publication are in scope of the Data Collection Policy.

Any prospective depositor who is unsure whether a dataset submission will be eligible for admission to the repository should contact researchdata@4tu.nl for advice.

4TU.ResearchData is responsible for the maintenance, review and revision of all its policies and documentation, including the Data Collection Policy.

If you have any comments or questions regarding this policy, please contact us at researchdata@4tu.nl.

Appendix: Subjects (not exhaustive)

Science and technology

- Mathematics
- Physics
- Chemistry
- Technology/Technical sciences
  - Materials science
  - Mechanical engineering, aerospace engineering
  - Electrical engineering
  - Civil engineering, building technology
  - Architectural engineering
  - Chemical engineering, process technology
  - Geotechnics
  - Industrial design engineering
  - Energy supply
  - Technology assessment
- Nanotechnology
- Biotechnology
- Earth sciences
- Computer science
- Astronomy, astrophysics
- Agriculture and physical environment
  - Exploitation and management of the physical environment
  - Plant production and animal production

Life sciences and medicine

- Life sciences
- Biology
- Medicine (human and animal)
  - Pathology, pathological anatomy
  - Organs and organ systems
  - Medical specialisms
- Health sciences
- Kinesiology
- Veterinary medicine