Data Collection Policy

1. Mission

4TU.Centre for Research Data (short: 4TU.ResearchData) is a data archive for technical-scientific research data that stores data in a permanent and sustainable manner, according to the guidelines of the international Data Seal of Approval.

The Data Seal of Approval (DSA), which provides the entry level to the European Framework, offers 16 quality guidelines related to trustworthy data management and stewardship.
Being a Trusted Digital Repository, we are demonstrating to researchers that we take appropriate measures to ensure the long-term availability and quality of data we hold.

Our mission is to ensure the accessibility of technical scientific research data during and after completion of research to give a quality boost to contemporary and future research.

4TU.ResearchData is currently a consortium of three partners: TU Delft, TU Eindhoven and Twente University.[1] The consortium is open to organizations that share, and actively contribute to, our mission.

2. Scope

The 4TU.ResearchData archive is primarily aiming at data from all fields and subjects in science and engineering but covers also subjects in the life sciences (see the appendix for a nominal list of subjects)

Atmospheric and environmental research is one of the interests 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 an 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 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 size of the datasets, our focus is on sharing of the richly diverse and heterogeneous small datasets, so-called long-tail data, as projects that produce large amounts of data ('Big Data') are the exception and often have subject-specific data services in place. Our focus is not store large datasets comprising petabytes of data, but to manage multiple data objects in a way that facilitates their reuse.

A dataset is defined 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.

---

[1] The larger 4TU Federation also includes Wageningen University & Research

Datasets may be deposited in the data archive 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 useable 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 Plan.

4. Removal of datasets

Once it has been published, 4TU.ResearchData prefers not to remove a dataset with a DOI, as it may have been cited by other researchers.
However, if serious grounds exists, for example in case of data falsification or even fabrication, 4TU.ResearchData can remove the dataset from the archive wholly or in part, or to restrict or prevent access to the dataset on a temporary or permanent basis. The depositor shall be informed in such cases.

In the unlikely event of a dataset being removed from public view, a landing page is still accessible from the DOI and the metadata associated with the DOI remains publicly available.

5. Licensing

When data is published, a licence will be attached to it to tell users how they may use the data. Currently, 4TU.ResearchData is providing a [bespoke license](#) in which is specified that, when re-using data, end users should clearly state the name(s) of the original author(s)/creator(s) and that they will not use the data for commercial purposes.

As there are now standard licences in place, we are considering the option that allows depositors to select a standard licence from a pre-defined list. This will increase the flexibility for depositors in choosing the appropriate licence for their data and is for end users easier to understand.

6. Versioning

A new version of a dataset should be created when an existing dataset is reprocessed, corrected or appended with additional data.

When a new version of a dataset is published, a new metadata landing page is created and a new DOI is minted. This way, the already existing persistent identifier (DOI) will continue to refer uniquely to the earlier version of the dataset. The new and the previous dataset are cross-referenced in their respective descriptive metadata.

Alternatively, when there is a minor change (e.g. correction of misspelling), this change is documented in the administrative metadata; no new persistent identifier is minted.

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 anonymized so that individuals, organisations or businesses cannot be identified.

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 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	Restricted access; only embargo period can be applied to provide	MEDIUM

		completed or finished. Data with commercial potential. Data resulting from sponsored or collaborative research.	delayed access to the data.	
Sensitive personal data	Data that can be traced back to an individual and that, if disclosed, could result in harm to that person	Human research data. Individually identifiable financial or medical information. Individually identifiable genetic information.	Not accessible through the internet. Currently not supported by 4TU.ResearchData.	HIGH

4TU.ResearchData is currently reviewing its restricted access functionality in order to provide better support for confidential data.

8 . Reuse

Since 2010 the data archive of 4TU.Centre for Research Data (previously known as 3TU.Datacentrum) 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 archive of its type in The Netherlands.

As of April 2017 we host >7200 datasets. According to our registration records we now have over 2100 users from 90 countries.

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

To encourage users to cite 4TU.ResearchData datasets, we provide a ready-to-go citation for each dataset that authors can use.

However, in practice, many authors use partial or non-standard citations, or fail to cite data at all. This makes measuring the reuse of data by tracking the citations very difficult.

Currently, 4TU.ResearchData measures the uploads, downloads of the data, page views, resolutions to the Digital Object Identifier (DOI), registered users, and unique logons.

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 archive, we will do all we can, and are eager to get involved, in establishing services that support these goals.

9. Charges

All our datasets can be downloaded free of charge for non-commercial purposes.
Every researcher, both in the Netherlands and abroad, can upload up to 10 GB of data per year to our data archive free of charge. For depositing additional data, there is a one-off cost of € 4.50 per GB, but 4TU staff are happy to discuss different possibilities for larger collections of research data.

By default, this data will be available free of charge for non-commercial purposes and be stored for a minimum of 15 years.

Researchers from partners of 4TU.Centre for Research Data (TU Delft, TU Eindhoven and University of Twente) are allowed to deposit up to 100 GB per year free of charge.

Where payment is required, invoices are sent by post or made available online.

10. Responsibilities

4TU.ResearchData staff will determine whether datasets submitted to the archive for deposit 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 archive 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