Publishing requirements for data and code for TU Delft PhD candidates

Whenever in doubt about the publishing requirements for data and code for TU Delft PhD candidates and/or if you need guidance about best practices for publishing data and code in general, please get in touch with your Faculty Data Steward.

What are the TU Delft Research Data Framework Policy requirements about data and code publication?

For PhD candidates: all PhD candidates who started on or after 1 January 2019 to deposit research data (and code) supporting their theses in a research data repository before they can graduate (unless there is a valid reason why this is not possible).

For PhD supervisors: Ensuring that PhD candidates make all data and code underlying their completed PhD theses FAIR (Findable, Accessible, Interoperable and Reusable) by sharing in a research data repository, which guarantees that data will be available for at least 10 years from the end of the research project, unless there are valid reasons which make research data unsuitable for sharing. (For all PhDs who started on or after 1 January 2019).

What research data and code refers to the TU Delft Research Data Framework Policy?

As research data and code, the policy refers to all research outputs necessary to validate and reuse the results presented in the thesis.

Validate means showing that your research outcomes are based on strong scientific evidence. A practical way of thinking about it is: what someone would need, for example a peer reviewer or other researcher, to trust your results and to reuse them?

A non-exhaustive list of examples include:

- References to re-used data and code
- Protocols/settings followed to generate or collect raw data
- Raw data

- Derived or intermediate data
- Finalised data
- Code to process the raw data
- Code developed as main research output, and the respective documentation
- Documentation about licensed software used to process the raw data

When is the right moment to publish research data and code?

(Ideally) With each scientific publication

- Relevant information underpinning the publication
- Some repositories (such as 4TU.ResearchData) allow you to reserve a DOI (Digital Object Identifier) of the data/code that you can add as a reference in your scientific publication
- After the publication is accepted, you make the data/code publicly available

At the end of the PhD project

- Data supporting unpublished chapters
- Other data/code that were unpublished so far (if any)
- Do not leave the documentation process until the end!

Remark:

Ideally, research data and code should be made publicly available at the same moment when you publish your research articles. Please note that you can also make data and code publicly available if you decide to publish preprints.

However, during a project (or PhD thesis) there might be research data, datasets, research software/code or other research outputs that you feel is not quite ready yet to make it publicly available and you would like to wait until you finalise the project (or PhD thesis).

That is fine BUT, make sure that you work on the documentation of those research outputs along the project and don't leave it until the end! Otherwise, leaving the preparation of the dataset or code and its documentation for publication until the end of the project (or end of the PhD thesis) might create a lot of delays on the finalisation of the project.

What research data and code (research outputs) not to publish?

In principle, for most research done at TU Delft, it is suitable to make the data and software/code underpinning research findings available in a data repository.

Valid reasons for not publishing research data and software/code could be:

- working with confidential data and/or software
- working/collecting personal data that cannot be anonymized or pseudonymized
- working/collecting data or developing software that can be severely misused or falls under special regulations, for example, export control
- data, software, research outputs that you do not have ownership on

Whenever in doubt, contact your Faculty Data Steward.

Definition of Personal data - all information about an identified or identifiable natural person (the data subject). A person is considered identifiable if they can be identified directly or indirectly based on one or more items of personal data, for example, name and address, ethnicity, date of birth and IP-address. In general, it can be assumed that personal data include all data relating to a living person that makes it possible to identify this person or to distinguish them uniquely from other persons.

A few examples of confidential data:

- national security data (for example nuclear research)
- data falling under export control regulations
- confidential data received from commercial, or other external partners
- data related to competitive advantage (such as patent, IP)
- data which could lead to reputation/brand damage (such as animal research, climate change, personal data)
- politically-sensitive data (such as research commissioned by public authorities, research in social issues)

What should happen to unpublished research data and code (research outputs)?

In summary:

- Data/Software/Code can be published. 4TU ResearchData is a good option for this

- Data/Software/Code can be internally archived. Project Data (U:) Drive/Staff -Umbrella is the recommended solution. The access to the drive should be managed (or transferred before the PhD candidate leaves) by a TU Delft employee (TUD supervisor, promotor, group leader), and this person should know how to access the drive.

- Data/Software/Code can be deleted. If data/code is irrelevant or too sensitive/confidential to be safely and legally preserved.

How to select data and code for publishing and/or archiving?

Publish/archive data and code that:

- is needed to **verify findings and protocols, and that allows others to build upon on your research**

- is needed to **replicate your results** - same analysis performed on different datasets produces qualitatively similar answers (relevant for those working with simulations)

- is of a **unique nature** such as is based on non-repeatable or costly observations

You should also:

Weigh up the **costs** between collecting the data again versus making the data FAIR and publishing/archiving them.

How to select a data repository?

Essential

- Be recognised in the research community

- Have clear terms and conditions

- Use common metadata standards for the dataset

- Provide persistent and unique identifiers (DOI/handle/...)

- Offer standard licences for data and/or code

- Can store data for at least 10 years

Optional

- Offer embargo periods and control over data access
- Enable dataset reviews
- Deliver download/citation statistics

4TU.ResearchData, Zenodo, DANS, Figshare are all trusted repositories that publish data according to the FAIR (Findable, Accessible, Interoperable, Reusable) principles by:

- making research data accessible, discoverable and available for the long term,
- providing persistent and unique identifiers like DOI to make data findable and citable
- offering standard licences that determines terms and conditions regarding sharing and reuse
- using common metadata standards to help others identify and discover the data.

If there are other disciplinary or domain repositories that are commonly used and endorsed by your research community, those might be more suitable to publish the data resulting from the project.

If you work with software/code, making it publicly available on GitHub/GitLab is not enough to comply with the TU Delft Research Software Policy. You need to register the software/code on 4TU.ResearchData or, if you are using another data repository, you need to register the DOI in TU Delft PURE.

What to check before publishing research data and code?

Data file organisation

- Use consistent and informative file names
- Proper folder structure:
  - Data, methods, and outputs should be clearly separated;
  - Store the raw data separated from the processed data;
  - The computational environment should be specified.

Data file quality

- The files can be opened (that is, not corrupted)
- The file format is open (that is, not proprietary)

- Recommended file formats : https://data.4tu.nl/s/documents/Preferred_File_Formats_2023.pdf
- The selected file format is recommended for data sharing, reuse and preservation.

Data documentation

- README file:
  - Write it in an open format such as .txt or .md (Markdown)
  - Make it clear what the README file is documenting
  - Structure it with defined sections:
    - General information
    - Methodological information
    - Sharing and access information
    - For code: include information on how to run the code!
  - Some good templates of README files are found here:
    - Data README (4TU.ResearchData)
    - Data / Code / ML models READMEs (TU Delft AE)

Additional Data documentation (if applicable)

- Codebooks (qualitative data)
- Data Dictionary (description of variables)
- Electronic Lab Notebooks (ELNs)
- Jupyter Notebooks (containing executable code, code outputs, (formatted) narrative text, formulas, etc.)
- Metadata files with additional (discipline-specific) metadata in an open or machine-readable file format

Software/Code documentation

README file:

- The goal of the project

- Installation instructions
- How can people get the software/code? Are there system/software requirements? What versions of packages, etc. were used?
- Licence information
- Citation information
- Optional: Issue reporting & contributing guidelines
- Video Utrecht University ‘Best Practices in Writing Reproducible Code- - ReadMe file for software (2 min)
- README example: Software / Code README (TU Delft DCC)

Licensing

4TU.ResearchData Licensing information

Open Software Licences: https://choosealicense.com/

Check requirements and guidelines about software licensing in the TU Delft Research Software Policy and the ‘Guidelines on Research Software Licensing, Registration and Commercialisation at TU Delft