

# Publishing Research Data & Code

---

Graduate School Information Session

Presenters: Deepshikha Purwar, Lora Armstrong, Nicolas Dintzner,  
Yasemin Türkyilmaz-van der Velden

Slides available at: <https://doi.org/10.5281/zenodo.10649047>



# Session outline

---

- Presentation (50 min)
- Q&A and closing (40 min)

# Who are we?

---

- Data Stewards Team
- 1 – 2 Data Stewards per faculty
- Support, train and encourage – best practices in research data & software management
- [Contact page](#)



Lora - CEG  
Data Steward



Nicolas - TPM  
Data Steward



Deepshikha - Library  
Data Steward



Yasemin - Data  
Stewardship Coordinator

# House rules

---

- This presentation is being recorded. It will be shared with participants in the session
- Feel free to ask questions in the Teams chat throughout the presentation
- The moderator will collect your questions throughout the presentation and they will be answered at the end

Why are we talking about  
data & code publication?

---

# Paradigm Shift

---

- Transparency and reproducibility boost **trustworthiness**
- Articles with linked data have up to 25% higher citation **impact** ([Colavizza et al., 2020](#))
- Saving time and resources increases **efficiency** and accelerates **innovation**
- Funder, institution and journal **requirements**



*“As open as possible, as closed as necessary”*

European Commission, 2016

# TU Delft & Faculty Policies

This page contains the general TU Delft Research Data Framework Policy and Research Software Policies and Guidelines Documents on the right, and faculty specific Research Data Management Policies below.

Please contact [your faculty data steward](#) for support or questions about the University and Faculty Policies and their implications for your work.



Policy  
requirement at  
TU Delft

<https://www.tudelft.nl/en/library/research-data-management/r/policies/tu-delft-faculty-policies>

Faculty of Applied Sciences  
Research Data Management Policy



AS

Faculty of Civil Engineering and Geosciences  
Research Data Management Policy



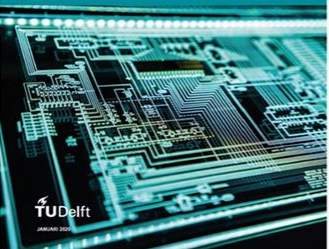
CEG

Faculty of Industrial Design Engineering  
Research Data Management Policy



IDE

Faculty of Electrical Engineering Mathematics and Computer Science  
Research Data Management Policy



EEMCS

Faculty of Technology, Policy and Management  
Research Data Management Policy



TPM

QuTech  
Research Data Management Policy



QuTech

Faculty of 3mE  
Research Data Management Policy



3mE

Faculty of Architecture and the Built Environment  
Research Data Management Policy



ABE

Faculty of Aerospace Engineering  
Research Data Management Policy



AE



# Research Data Policy requirement at TU Delft

---



## **In each Faculty Policy, PHD CANDIDATES are responsible for:**

- *Developing a written data management plan for managing research outputs within the first 12 months of the PhD study (For PhD candidates who started on or after 1 January 2020).*
- *Attending the relevant training in data management.*
- *Ensuring that all data and code underlying completed PhD theses are 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).*

# Research Data Policy requirement at TU Delft

---



## In each Faculty Policy, PHD SUPERVISORS are responsible for:

- *Supporting PhD candidates in preparation of a written data management plan for managing research outputs within the first 12 months of their PhD. (For all PhD candidates who started on or after 1 January 2020).*
- *Ensuring that PhD candidates attend relevant training on data management.*
- *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).*

# Research Software Policy at TU Delft

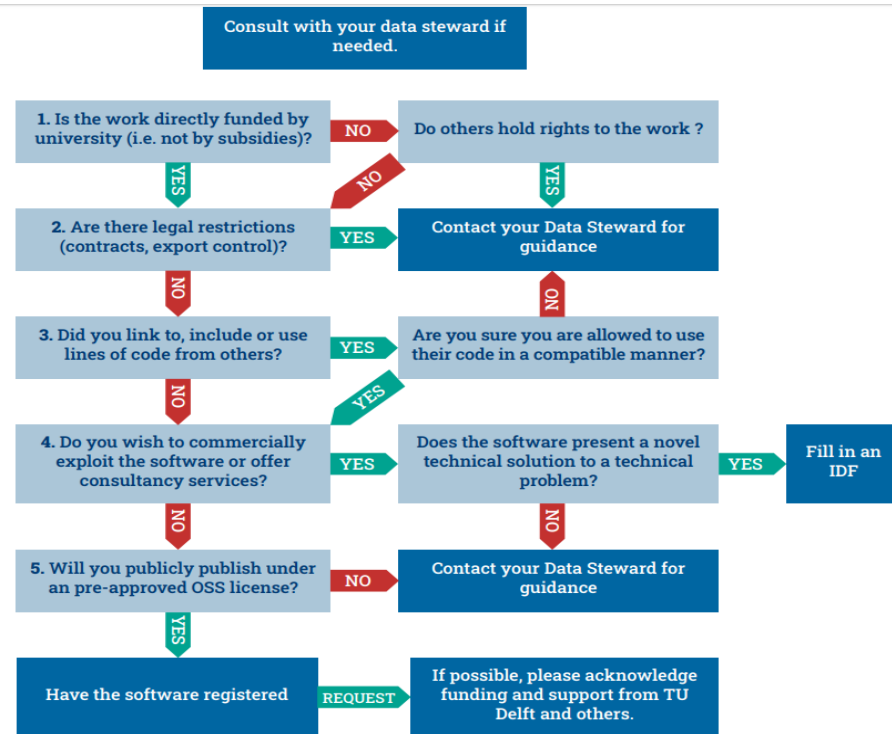


Figure 1. Decision tree to guide software developers, researchers and staff on when they can apply an open source licence to their software. OSS: Open Source Software, IDF: Invention Disclosure Form

Decision Tree to publish software as open source where possible and appropriate by considering:

- the possibility of commercial exploitation
- the risks of Intellectual Property Rights (IPR) infringement
- compatibility of licenses

For more info:

[slide deck](#) and [recording of an information session](#).

[Guidance page](#)

# Fulfilling the data/code publishing requirement

---

## Guidance for doctoral candidates completing their studies

Well done! You are about to complete your PhD studies at TU Delft. Before you do this and leave TU Delft, you need to publish your research data and code, so that others can reproduce your findings and re-use your research outcomes. In accordance with the [TU Delft Research Data Framework Policy](#), all doctoral candidates who started on, or after 1 January 2019, are required to upload their research data to [4TU Centre for Research Data](#) (or another suitable data archive) before their defence ceremony.

PhD candidates and their supervisors are advised to consult the [Research Data Archiving Checklist](#) before signing [Form B](#)

All doctoral candidates who started before 1 January 2019 are strongly encouraged to upload their data.

## What exactly do I need to do?

You need to:

1. [Gather and organise the research data \(and code\) which support the results described in your PhD thesis](#)
2. [Ensure that your data is not confidential](#)
3. [Describe your data and code](#)
4. [Upload your research data \(and code\) and the description of your data to 4TU Centre for Research Data](#)
5. [Add the DOI hyperlink to your data \(and code\) into your thesis](#)
6. [Submit your thesis for examination](#)

- [Guidance page](#)
- Research Data Archiving Checklist
- Step-by-step guide

# Expected outcome of this session

---

- Understand the paradigm shift and policy requirement
- Clarify the definition of data and publishing
- Know what data to publish
- Know how to prepare the data for publishing
- Know how to select a repository
- Know all the available support

# Definitions

---

# First...what 'data' are we talking about?

---



Image by [mmi9](#) from [Pixabay](#)



Image by [OpenClipart-Vectors](#) from [Pixabay](#)

All research outputs necessary to **validate and reuse the results presented in the thesis**

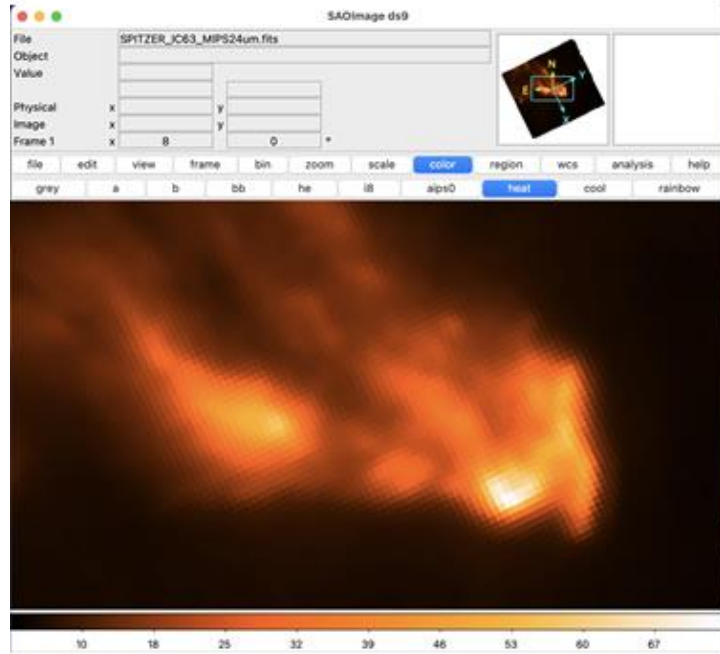
# Examples

---

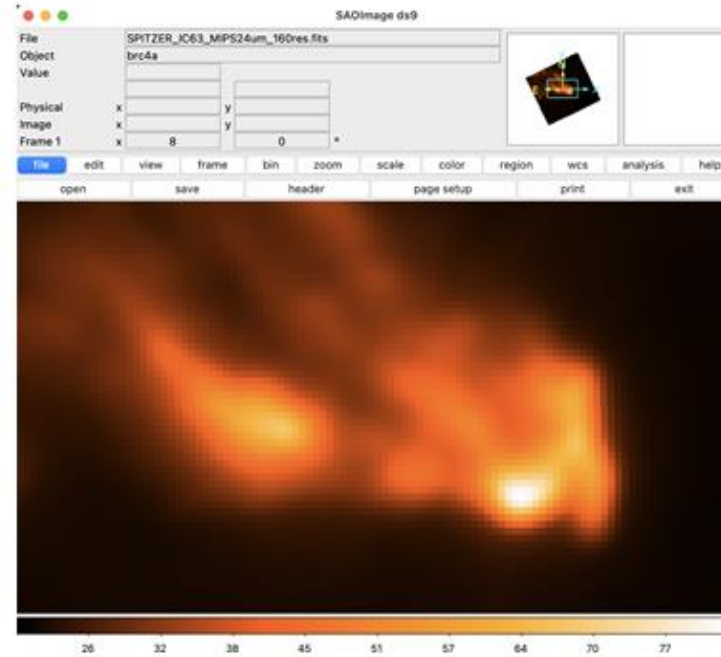
- References to **re-used** data and code
- Protocols/settings followed to **generate** or **collect** raw data
- **Raw** data
- **Derived or intermediate** data
- **Finalized** data
- **Code to process** the raw data
- **Code developed** as main research output, and the respective documentation
- **Documentation about licensed software** used to process the raw data



## Raw data



## Processed data #1



## Processed data #2

#wave	#flux_density	#unc_flux_density
#um	#MJy/sr	#MJy/sr
3.6	1.31e-06	3.97e-08
4.5	4.44e-07	1.33e-08
5.8	3.44e-06	1.21e-07
8.0	6.95e-06	2.10e-07
24.0	3.54e-06	3.62e-07
70.0	1.45e-05	2.18e-06

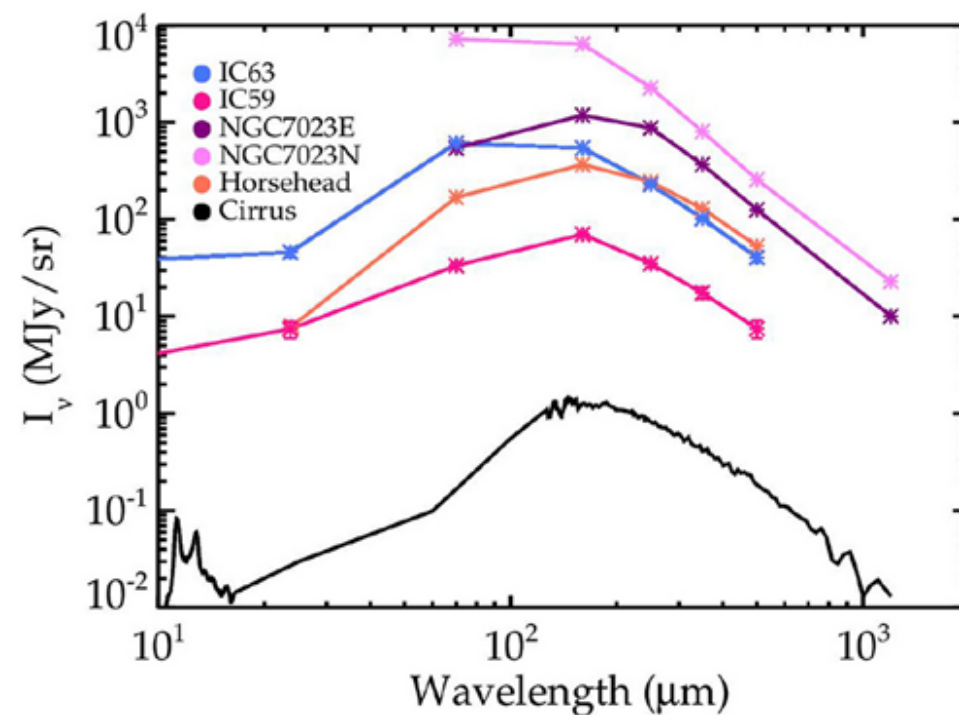
Accompanied by documentation (e.g. README file)

## Processed data #2

#wave	#flux_density	#unc_flux_density
#um	#MJy/sr	#MJy/sr
3.6	1.31e-06	3.97e-08
4.5	4.44e-07	1.33e-08
5.8	3.44e-06	1.21e-07
8.0	6.95e-06	2.10e-07
24.0	3.54e-06	3.62e-07
70.0	1.45e-05	2.18e-06

... #N processing steps ...

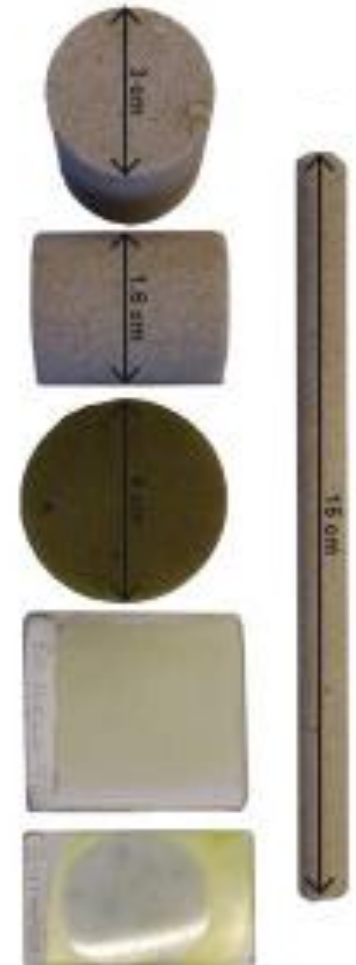
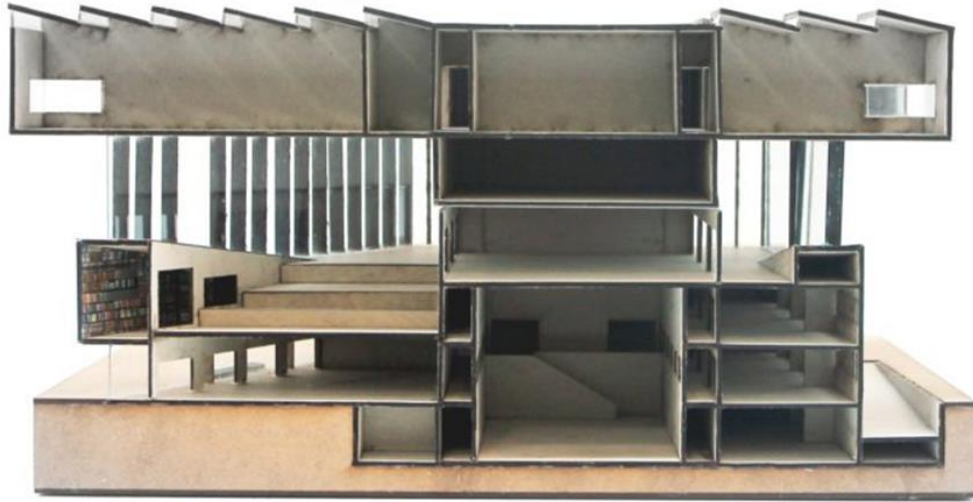
Finalized data



Accompanied by documentation (e.g. README file)

<https://doi.org/10.1051/0004-6361/201832808>

# Data?



<https://doi.org/10.1016/j.marpetgeo.2015.06.001>

# What does publishing mean?

## The Obstacle Detection and Avoidance Dataset for Drones

doi: [10.4121/14214236.v1](https://doi.org/10.4121/14214236.v1)

The data/code have a DOI (Digital Object Identifier).

Cite

### DATASET

by [Julien Dupeyroux](#), [Nikhil Wessendorp](#), [Raoul Dinaux](#), [Guido De Croon](#)

We introduce the Obstacle Detection and Avoidance Dataset for Drones, aiming at providing raw data obtained in a real indoor environment with sensors adapted for aerial robotics. Our micro air vehicle (MAV) is equipped with the following sensors: (i) an event-based camera, the dynamic performance of which make it optimized for drone applications; (ii) a standard RGB camera; (iii) a 24-GHz radar sensor to enhance multi-sensory solutions; and (iv) a 6-axes IMU. The ground truth position and attitude are provided by the OptiTrack motion capture system. The resulting dataset consists in more than 1350 samples obtained in four distinct conditions (one or two obstacles, full or dim light). It is intended for benchmarking algorithmic and neural solutions for obstacle detection and avoidance with UAVs, but also course estimation and therefore autonomous navigation. For further information, please visit: [https://github.com/tudelft/ODA\\_Dataset](https://github.com/tudelft/ODA_Dataset)

### HISTORY

2021-03-19 first online, published, posted

### PUBLISHER

4TU.ResearchData

### FUNDING

- Framework of key enabling technologies for safe and autonomous drones' applications (COMP4DRONES) (grant code 826610) [[more info...](#)]  
Ministry of Education and Science

### ORGANIZATIONS

TU Delft, Faculty of Aerospace Engineering

### DATA

Data/code are publicly available for download (unless valid reason to restrict access)

### FILES (1)

98,186,579,073 bytes MD5:189639db8176ccdbd728b88d99c27309 [Dupeyroux\\_et\\_al\\_2021\\_ODA\\_DATASET\\_Full.zip](#)

[download all files \(zip\)](#)

98,186,579,073 bytes unzipped



### USAGE STATS

**3321** **13225**

views downloads

### CATEGORIES

[Artificial Intelligence and Image Processing](#)  
[Aerospace Engineering](#)

### KEYWORDS

[event-based cameras](#), [Micro Air Vehicle](#), [drone](#), [neuromorphic camera](#), [neuromorphic sensors](#), [Radar sensor](#), [RGB camera](#), [Robot operating system \(ROS\)](#), [Unmanned Aerial Vehicle \(UAV\)](#)

### LICENCE



### EXPORT AS...

[RefWorks](#), [BibTeX](#), [Reference Manager](#), [Endnote](#), [DataCite](#), [NLN](#), [DC](#), [CFF](#)

The metadata (data about data) are publicly available.

<https://doi.org/10.4121/14214236>

# Things you should know before publishing

---

# When to publish

---

## **With each scientific publication**

- Relevant information for this paper
- After the publication is accepted

## **At the end of the PhD process**

- Data supporting unpublished chapters
- Other data/code that were unpublished so far (if any)

# What to publish

---

- Data & code needed to verify and/or reproduce your findings
- Data & code with a high potential for reuse

# What not to publish

---

- Confidential data
- Personal data that cannot be anonymized or pseudonymized
- Data that can be severely misused or falls under special regulations, e.g. export control

What happens to the unpublished data/code/?



# Off-boarding process for data/code

---

*All storage solutions provided by TUD are attached to an individual in nature. They will be deactivated and the data lost or hard to access once the contract with TUD is over. Therefore, before leaving TUD, one of the following should be done with data and code:*

## Published

- Data/code published in a research data repository (open or restricted access)
- 4TU.ResearchData is a good option

## Deleted

- If data/code is irrelevant or too sensitive/confidential to be safely and legally preserved

## Archived internally

- Archived in institutional storage: Project Data (U:) Drive/Staff -Umbrella (recommended)
- Access to data/code should be managed by TUD employee (TUD supervisor, promotor, group leader, etc.)
- Contact person should have knowledge about and know where to find the data/code

# Archiving data on the Project Data (U:) drive

HOME > RESEARCH SUPPORT > IT FOR RESEARCHERS > DATA STORAGE FOR RESEARCH: PROJECT DATA (U:)

## Data storage for Research: Project data (U:)

Data storage for research and, if needed accessible for externals

The **standard data storage** consists of:

- Storage space can include several terabytes. (Fair use)
- A backup is made on a daily basis and is stored for two weeks. This means a data loss of a maximum of one day can occur.

You can also choose to store the data for 14 days + 53 weeks

- Single storage location, not redundant. If the storage location fails, the data is temporarily unavailable. The back-up will then be reset.

With the help of CIFS and NFS (with kerberos authentication) you have access to the data.

- For a request of more than 5TB, your request is sent to the FIM.

Request Data Storage >

ICT malfunction or request ICT service >

Request storage using <https://tudelft.topdesk.net>

# Archiving data on the Project Data (U:) drive

---

## **If supervisor requests storage (best)**

- Give access to PhD by sending their netID to the service desk or in UMRA (<https://www.tudelft.nl/en/it-manuals/umra>)

## **If PhD requests storage**

- List supervisor as backup owner when request is made
- Transfer ownership to supervisor by sending their netID to the service desk

# The 'How to' for publishing data and code

---

# How to select data & code

---

You should publish/archive data and code that is:

- needed to **verify findings and protocols, and that allows others to build upon on your research** (funders and journals may require you to do this too)

# How to select data & code

---

Also consider data and code that are:

- Needed to **replicate your results** - *same* analysis performed on *different* datasets produces qualitatively similar answers (relevant for those working with simulations)
- of a **unique nature** e.g. is based on non-repeatable or costly observations

# How to select data & code

---

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

# Tips for different data types

---



# 1. Personal data

---

**Personal data is** information which can be used to **identify individuals**

General rule:

**personal data is never made publicly accessible**, unless clearly agreed on by the concerned individuals (i.e. through Informed Consent)

# Anonymized Personal data

---

**Anonymization** is the process of **removing personally identifiable information** from data sets.

If **absolutely no** relationships exists anywhere between the anonymized data and the people from whom the data was collected, the data can be publicly archived (i.e. published).

# Pseudonymized data

---

**Pseudonymization** is the process of de-identification by which personally identifiable information fields within a dataset are **replaced by one or more artificial identifiers, or pseudonyms**

**Pseudonymized data** must be published under restricted access, or archived in internal storage under responsibility of a specified role

## 2. Licensed/Commercial data & code

---

### Examples:

Publicly accessible data and software/code distributed under a specific license or 'Terms of Use':

- Social media data, pictures from the internet, data from NGO
- Software/code you re-used that are under a specific license

## 2. Licensed/Commercial data & code

---

### Examples:

Data from private company or industry project partners:

- Commercial/confidential in nature, access is granted for research purposes in the context of the project

## 2. Licensed/Commercial data & code

---

- Working with such data/code is normally not a problem.
- Redistribution and/or publication, can **only be done in compliance** with
  - terms of the assigned license
  - terms of use declared by the data provider
  - clauses established in a collaboration agreement.
- E.g.: you can do research using content from Twitter, but you are not allowed to publish the "tweets" content.

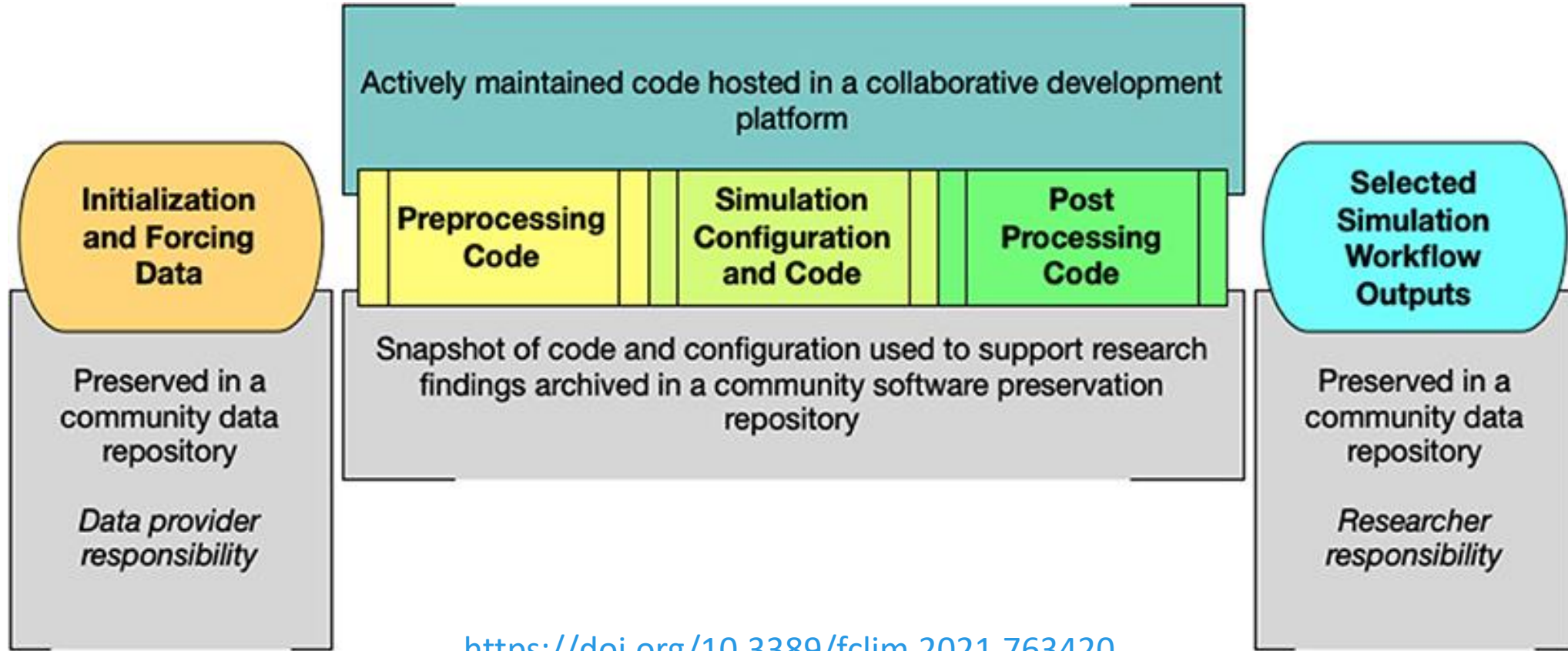
# 3. Numerical Simulations

---

- **Knowledge-production** vs. **data production** as a goal?
- Most research focuses on **knowledge production**
  - Typically share selected derived data outputs
  - Focus on workflows, smaller derived fields that communicate environmental state, other information important for building similar studies in future
- Studies focusing on **data production** will need to share much more data in addition to information about workflows, etc.

‘Open Science Expectations for Simulation-Based Research’: <https://doi.org/10.3389/fclim.2021.763420>  
Rubric for selecting data to preserve + project use cases: <https://doi.org/10.5065/g936-q118>

# 3. Numerical Simulations



<https://doi.org/10.3389/fclim.2021.763420>

Preserve and share **all elements of the simulation workflow**, not just model source code



# 4. Large datasets

---

- Several to many **TB** in size
- Ideally project has budgeted for data sharing costs (e.g. in [SURF data repository](#)) but if not...
- Consider if data volume can be **reduced or divided**
  - Can data be published in separate chunks?
  - Can data be compressed?
  - Can sampling, aggregation, binning, or other methods help?
  - Can data be regenerated & therefore not included?
  - Is sharing representative data enough for validation?

## 5. Samples- publishing & archiving

---

- **Document sample metadata** and which measurements have been made on each sample (publishing *or* archiving)
- Samples can be archived internally
  - Label samples with **ID number**, QR code or barcode
  - Document or make a map of your **sample storage location(s)**
  - **Designate** someone (supervisor, lab manager, etc) to be responsible for samples and be sure they know where to find them + associated data

# Research data repositories

---

# How to select a research data repository?

---

## *Essential*

- Be recognized 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

# How to select a research data repository?

---

## *Optional*

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

- Permanent and sustainable data repository: CoreTrustSeal
- Digital Object Identifier (DOI)
  - Can be assigned at every level of details: whole collection, each component within a collection
  - Can be reserved - e.g. to facilitate peer-review process of articles
- Data & software usage licence can be assigned

- Access level
  - Open access
  - Temporary embargoed access
  - Restricted access
  - Metadata-only record
  - Private link / URL
- Git connection
- Up to 1 TB per year free of charge
- One-time fee €1.50/GB for large datasets
- Data/code stored 15 years

- Resources

- The upload process

<https://data.4tu.nl/info/about-your-data/publish-cite>

- Uploading data

<https://data.4tu.nl/info/en/use/publish-cite/upload-your-data-in-our-data-repository>

- Data/code upload instructions

<https://www.youtube.com/watch?v=VmQr5oDXpoY>



What to check before  
publishing data/code?

---

# What to check before publishing data/code?

---

## Data file organization

- 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 (i.e. not corrupted)
- The file format is open (i.e. not proprietary)
- Recommended file format : [https://data.4tu.nl/s/documents/Preferred\\_File\\_Formats\\_2023.pdf](https://data.4tu.nl/s/documents/Preferred_File_Formats_2023.pdf)
  - The selected file format is recommended for data sharing, reuse and preservation.

# What to check before publishing data/code?

---

## Data documentation

[Guidelines for creating a README for data](#)

- README file:
  - Write it in an open format e.g. .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!

**Mode I fatigue delamination growth in composite laminates with fibre bridging**

Authors: L. Yao, R.C. Alderliesten

Affiliation: Structural Integrity & Composites Group, Faculty of Aerospace Engineering,  
Delft University of Technology

Corresponding author: R.C. Alderliesten

Contact information:

email:

address:

**General introduction**

This dataset contains data collected during crack growth experiments at Delft University of Technology, as part of Liaojun Yao's PhD Thesis project (December 2015):  
doi:10.4233/uuid:66e210e1-c884-45d6-b9d4-711907680452

**Test equipment**

All tests were performed on a 10 kN MTS fatigue test machine. The crack length was measured by means of a camera system.

The applied force and displacement were measured by the fatigue test machine, and also sent as inputs to the camera, in order to facilitate synchronisation of the data.

**Data organisation and naming**

The data included in this data set has been organised per specimen. The files follow the nomenclature system: Sp\_X\_Data\_analysis\_Y with

X = the specimen number 1 to 56

Y = indicating the number of runs with the same specimen.

# Example

General information, e.g. title, authors, and link to publication

Methodology information, e.g. test equipment

Other information, e.g. organisation and naming convention

# What to check before publishing data/code?

---

## 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

# What to check before publishing data/code?

---

## Code documentation

[2 min video about writing README for code](#)

- 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?
  - License information
  - Citation information
  - Optional: Issue reporting & Contributing guidelines

# CREATIVE COMMONS LICENSES







  
COPY & PUBLISH

  
ATTRIBUTION REQUIRED

  
COMMERCIAL USE

  
MODIFY & ADAPT

  
CHANGE LICENSE

	PUBLIC DOMAIN
	CC BY
	CC BY-SA
	CC BY-ND
	CC BY-NC
	CC BY-NC-SA
	CC BY-NC-ND

	COPY & PUBLISH	ATTRIBUTION REQUIRED	COMMERCIAL USE	MODIFY & ADAPT	CHANGE LICENSE
PUBLIC DOMAIN	✓	✗	✓	✓	✓
CC BY	✓	✓	✓	✓	✓
CC BY-SA	✓	✓	✓	✓	✗
CC BY-ND	✓	✓	✓	✗	✓
CC BY-NC	✓	✓	✗	✓	✓
CC BY-NC-SA	✓	✓	✗	✓	✗
CC BY-NC-ND	✓	✓	✗	✗	✓

  
You can redistribute (copy, publish, display, communicate, etc.)

  
You have to attribute the original work

  
You can use the work commercially

  
You can modify and adapt the original work

  
You can choose license type for your adaptations of the work.

# Licences for Data

[4TU.ResearchData Licensing information](http://4TU.ResearchData.com/licensing)

[Creative Commons licenses](#) by [Foter](#) (CC-BY-SA)

# Licences for Software

Original	CCO	MIT	BSD	Apache	EUPL	GPL, AGPL or LGPL	Proprietary
Combine with?							
CCO	YES	YES	YES	YES	YES	NO	NO
MIT	YES	YES	YES	YES	YES	NO	NO
BSD	YES	YES	YES	YES	YES	NO	NO
Apache	YES	YES	YES	YES	YES	NO	NO
EUPL	YES	YES	YES	YES	YES	NO	NO
GPL, AGPL or LGPL	YES	YES	YES	NO	NO	YES	NO
Proprietary	YES	YES	YES	Claused	NO	NO	Depends on licence

[4TU.ResearchData  
Licensing  
information](https://4TU.ResearchData.com/licensing-information)

<https://choosealicense.com/>

From: TU Delft Guidelines on Research Software



# Data & Code Publication Examples

---

## 4TU.ResearchData

"Qualitative coding of 12 semi-structure interviews on food behaviour context and food reporting engagement" (IDE)

- Data: <https://doi.org/10.4121/uuid:02b93c7c-545d-4501-b375-6db1aff039c6>

"Transport Patterns of Global Aviation NOx and their Short-term O3 Radiative Forcing – A Machine Learning Approach" (AE)

- Data: <https://doi.org/10.4121/16886977.v1>
- Figures: <https://doi.org/10.4121/20338212.v1>
- Article: <https://doi.org/10.5194/acp-2022-348>

# More examples

---

## 4TU.ResearchData

"Atmospheric Observations IDRA, Cabauw" (CEG)

- Collection of netCDF datasets: <https://doi.org/10.4121/collection:cabauw>

"Interviews about the educational interview tinkering with technology" (TPM)

- Personal data with restricted access: <https://doi.org/10.4121/uuid:02b93c7c-545d-4501-b375-6db1aff039c6>

"Code accompanying the paper "Validating human driver models for interaction-aware automated vehicle controllers: A human approach" (ME)

- Software publication with associated github repository: <https://doi.org/10.4121/16847203>

# Final check

---

- We want to do better than the current working practices
- We do not all have data
- Current practices may not "tick" all the boxes:
  - As long as the data / code are published, and the means of publication are deemed reasonable by the supervisory team, it is fine
  - As long as the quality of the data /code is "sufficient" for the supervisory team, it is fine
- Not happy with the published data / code?
  - In the 4TU.ResearchData repository you can create a new version (same DOI)

# Available support

---

- [Faculty Data Stewards](#)
- 4TU.ResearchData [researchdata@4tu.nl](mailto:researchdata@4tu.nl)
- Other relevant resources:
  - The TUD Library website: <https://www.tudelft.nl/en/library/research-data-management/r/publish/publish-research-data>
  - Copyright team from the library : <https://www.tudelft.nl/library/copyright>
  - Privacy team (personal data): [privacy-tud@tudelft.nl](mailto:privacy-tud@tudelft.nl)
  - Anything else? Not sure who to contact ? Check with your faculty data steward.

# Feedback and suggestions

---

Survey: <https://evasys-survey.tudelft.nl/evasys/online.php?p=E5DHK>

Short URL: <https://edu.nl/urv48>

QR Code:



[edu.nl/urv48](https://edu.nl/urv48)



Q & A