Publishing/Archiving Data & Code

Graduate School Information Session

Presenters: Paula Martinez Lavanchy, Nicolas Dintzner, Yan Wang

Slides available at: https://doi.org/10.5281/zenodo.7674963

Session outline

- Presentation (40min)
- Q&A and closing (20min)

House rules

- Please turn off your camera during the presentation
- Feel free to ask questions in the Q&A in Teams throughout the presentation
- You can turn on your camera when asking questions after the presentation

Why are we talking about data & code publication?

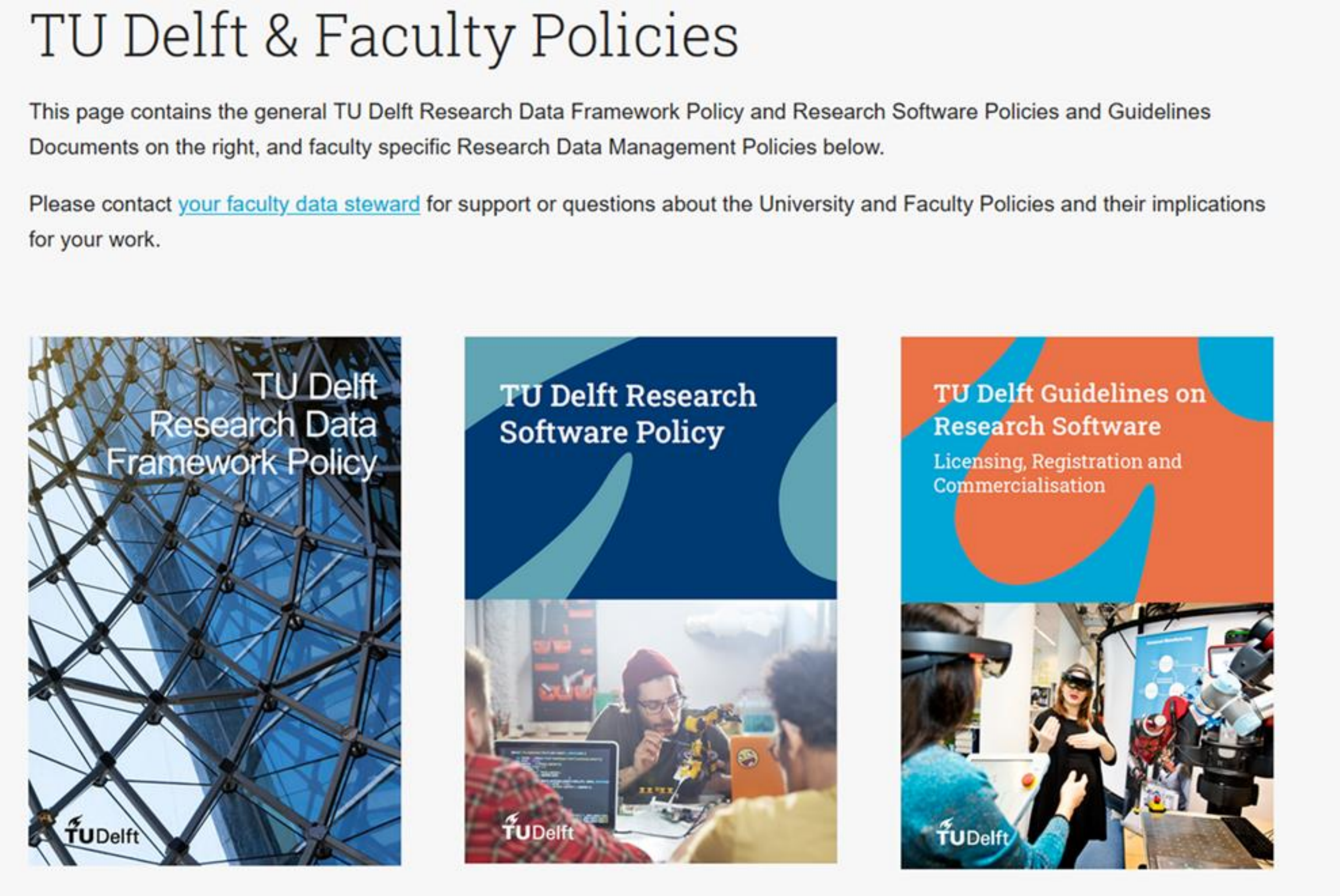
Paradigm Shift

- 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

Policy requirement at TU Delft



TU Delft & Faculty Policies

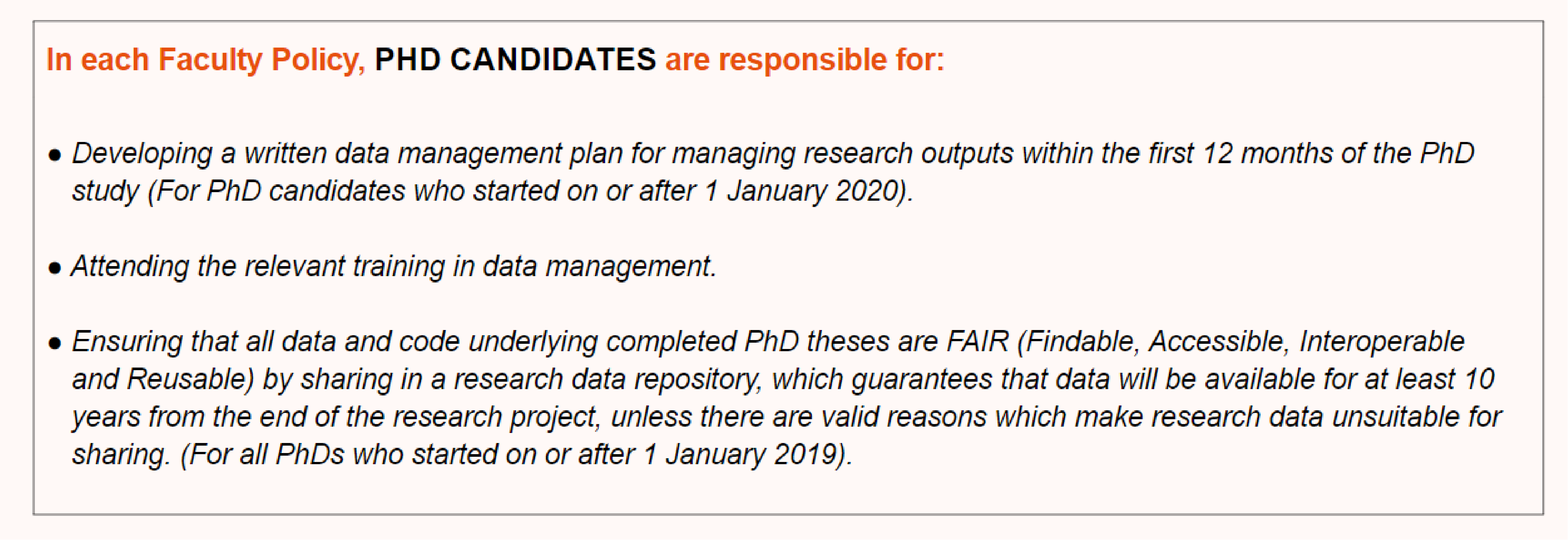
This page contains the general TU Delft Research Data Framework Policy and Research Software Policies and Guidelines Documents on the right, and faculty specific Research Data Management Policies below.

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

TU Delft Research Data Framework Policy
TUDelft
TU Delft Research Software Policy
TUDelft
TU Delft Guidelines on Research Software
Licensing, Registration and Commercialisation
TUDelft

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

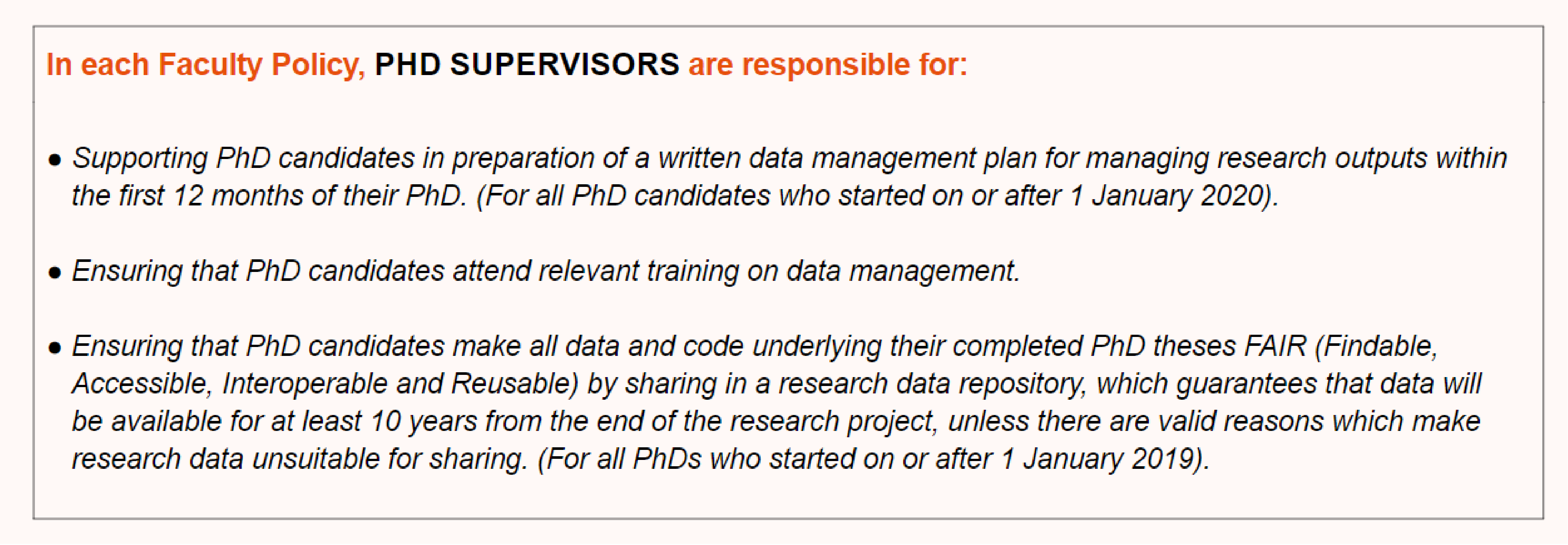
Policy requirement at TU Delft



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).*

Policy requirement at TU Delft

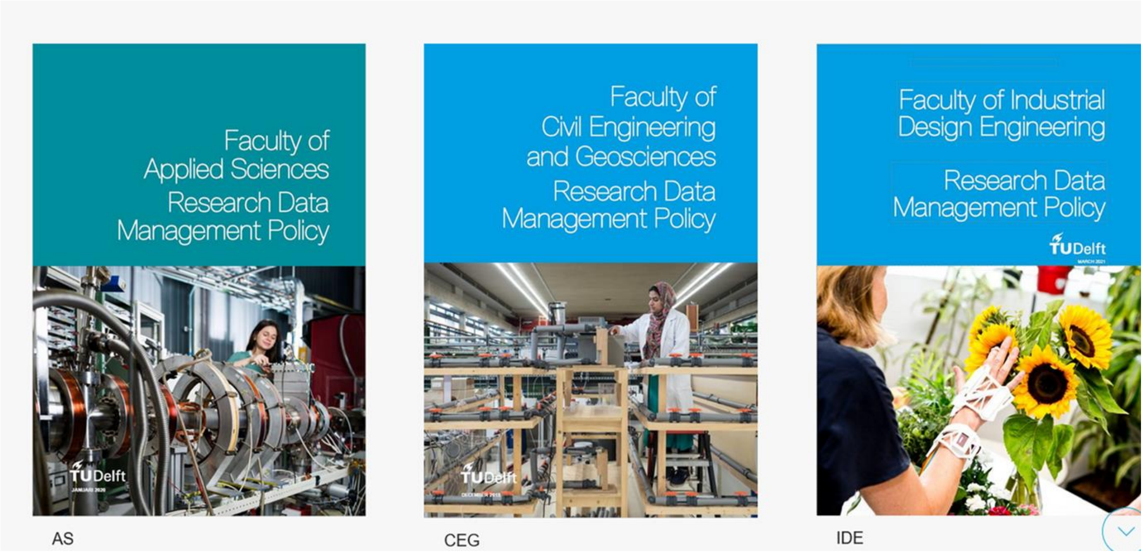


In each Faculty Policy, PHD SUPERVISORS are responsible for:

- *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).*



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

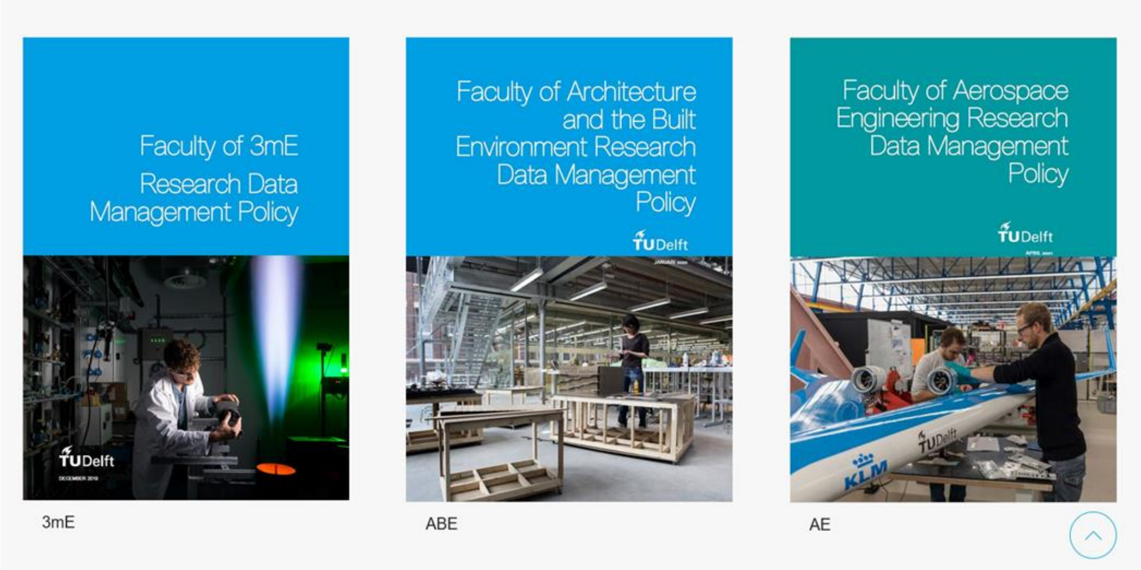
EEMCS

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

Menti

www.menti.com

Code: 38804337

Expected outcome of this session

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

Definitions

First…what ‘data’ are we talking about here?

.ipynb
.mat
.txt
.py
.c
.f90
.dat
…
…
.rpt
.plt
…

Image by mmi9 from Pixabay

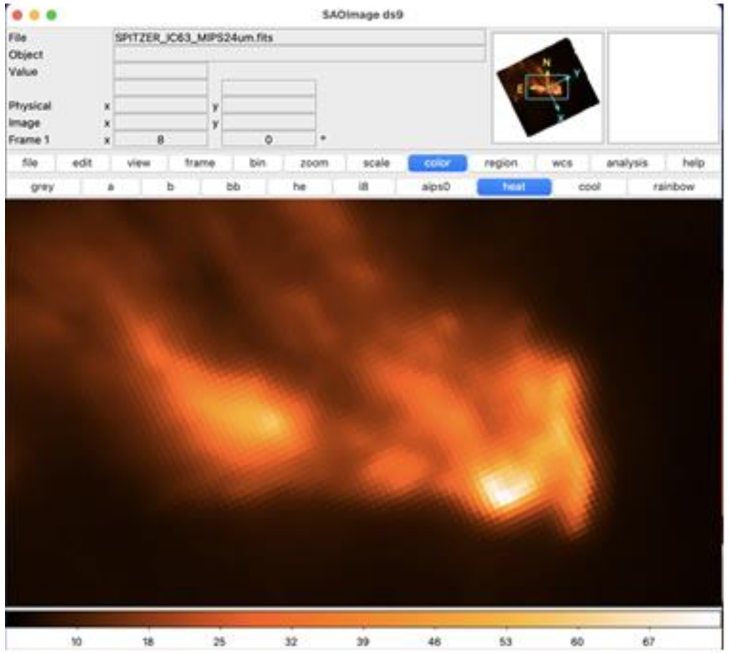
Image by OpenClipart-Vectors from Pixabay

All research output 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
- **Code to process** the raw data
- **Documentation about licensed software** used to process the raw data
- **Code developed** as main research output, and the respective documentation
- **Derived or inter-mediate** data
- **Finalized** data

Raw data

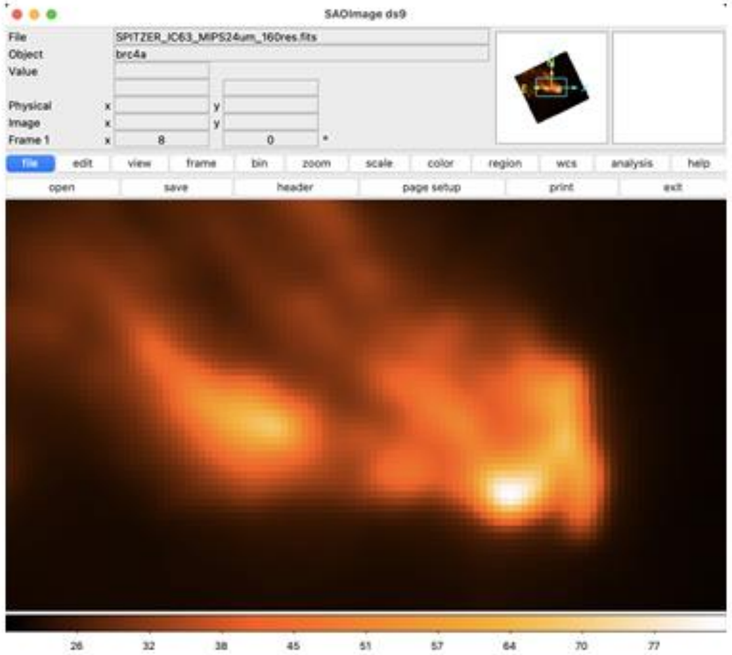


File SPITZER_IC63_MIPS24um.fits

Object

Frame 1 x 8 0

Processed data #1

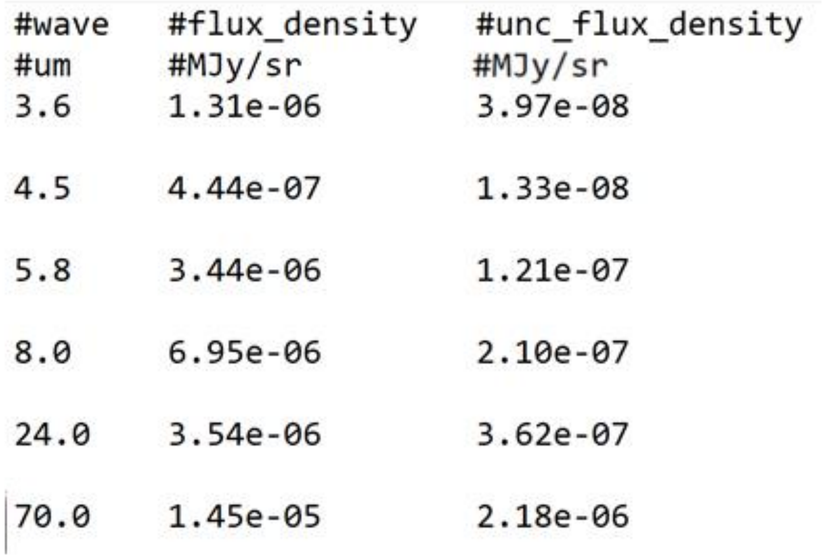


File SPITZER_IC63_MIPS24um_160res.fits

Object brc4a

Frame 1 x 8 0

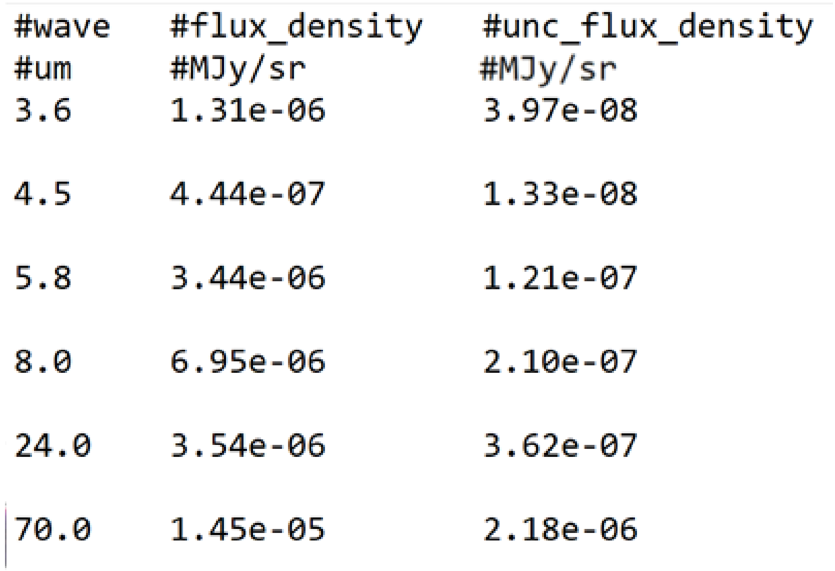
Processed data #2



```
#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
```

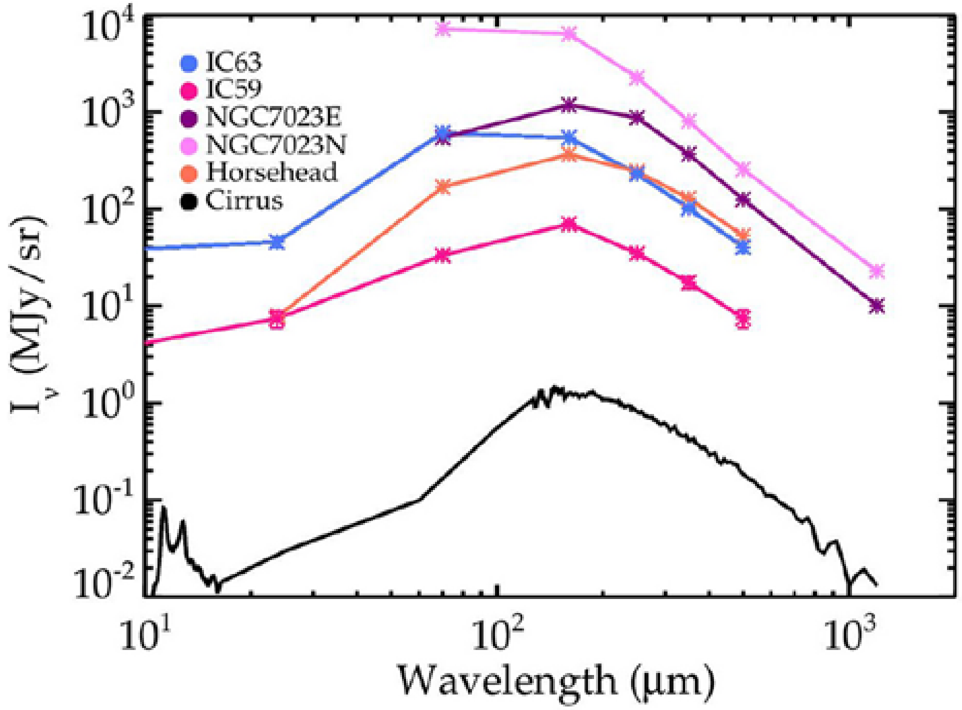
Accompanied by documentation (e.g. README file)

Processed data #2 ... #N processing steps ... Finalized data



```
#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
```

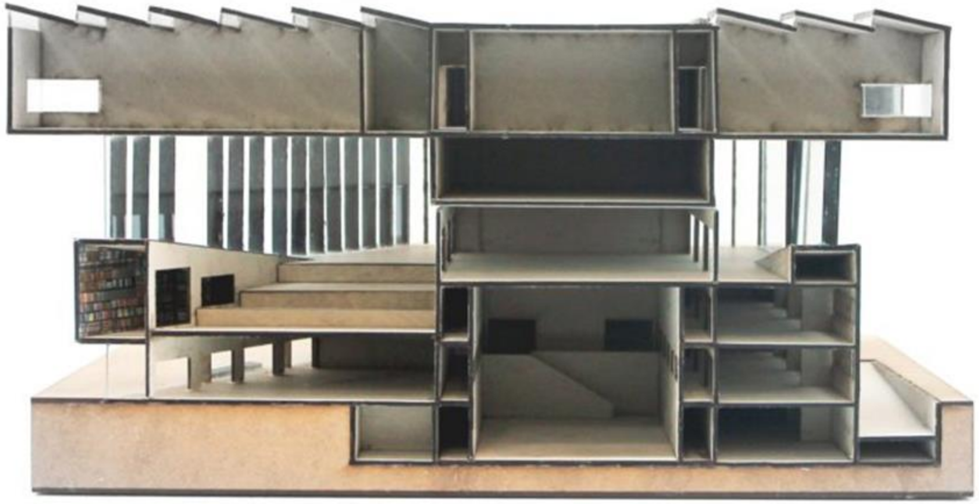
...



Accompanied by documentation (e.g. README file)

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

Data ?



Publishing vs Archiving

Publishing:

- The dataset is publicly visible (has a DOI, public metadata)
- The dataset is accessible to all or not (closed vs open datasets)

Archiving:

- The dataset is preserved "internally" (i.e. Institutional servers) for an expected long period of time
- Not storage, need to be FAIR (e.g. metadata, documentation)
- Contact / responsible person

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 was 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 anonymised or pseudonymised
- Data that can be severely misused or 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, data and code must be:

Publicly archived

- Data/code published in an open data repository
- 4TU.ResearchData is a good option

Deleted

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

Closed Archiving

- Archived in a research data repository with restricted access
- 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 where to find the data/code

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)

Information adapted form "Five steps to decide what data to keep" from the Digital Curation Centre and the Essentials 4 Data Support course chapter "Data selection" from RDNL

How to select data & code

Consider as well data and code that is:

- 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

Information adapted form “Five steps to decide what data to keep” from the Digital Curation Centre and the Essentials 4 Data Support course chapter “Data selection” from RDNL

How to select data & code

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

Information adapted form "Five steps to decide what data to keep" from the Digital Curation Centre and the Essentials 4 Data Support course chapter "Data selection" from RDNL

Personal data

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

General rule:

personal data is **never made publicly accessible**, unless clearly stated by concerned individuals (i.e. 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 who the data was collected from, 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

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

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

Licensed/Commercial data & code

- Working with such data/code is normally not a problem.
- Redistribution and/or publication, can only be done 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.

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

How to select a research data repository?

Optional

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

4TU.ResearchData

SCIENCE • ENGINEERING • DESIGN

CORE TRUST SEAL

- 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

4TU.ResearchData

SCIENCE • ENGINEERING • DESIGN

CORE TRUST SEAL

- 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

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

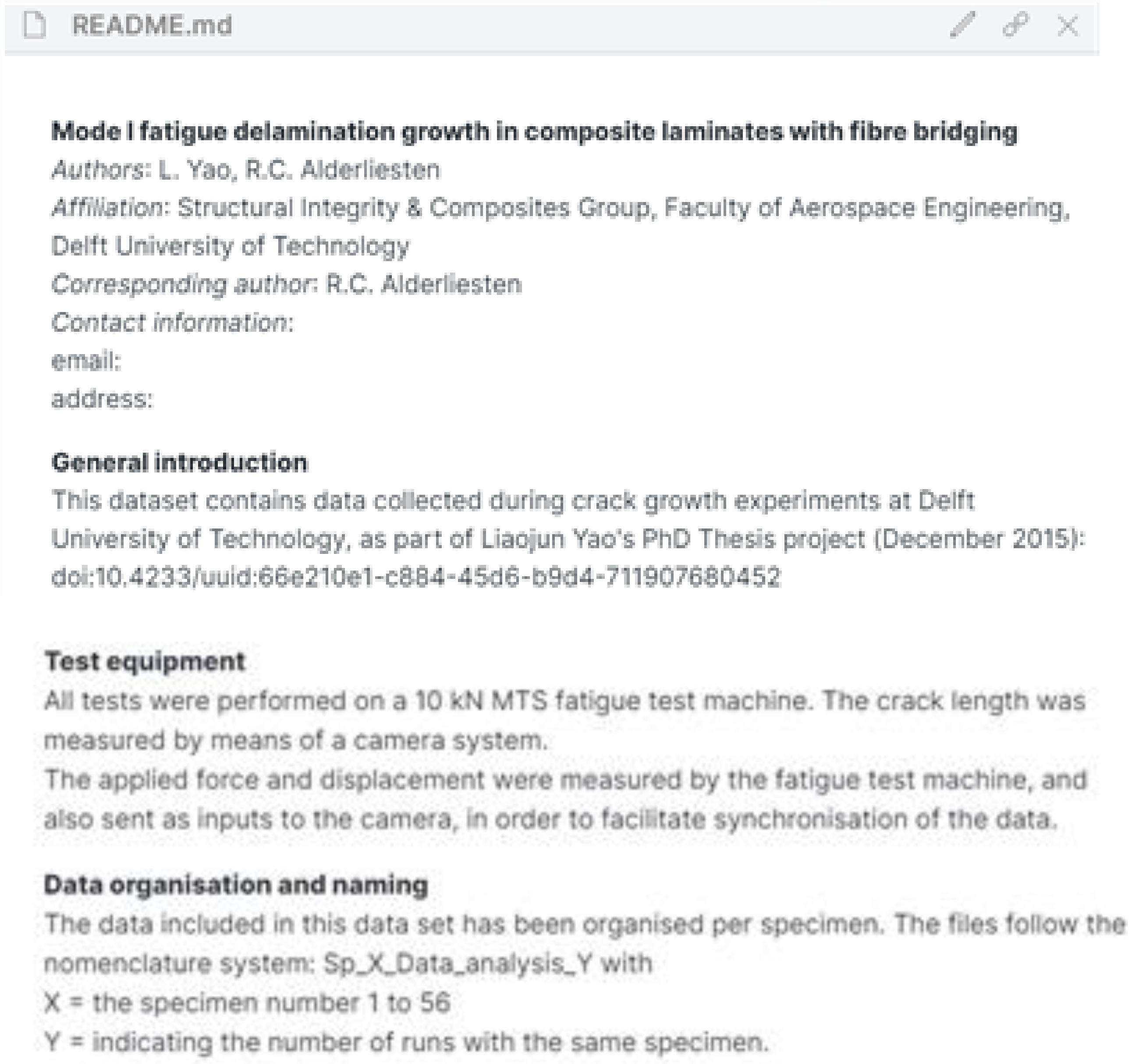
- The files can be open (i.e. not corrupted)
- The file format is open (i.e. not proprietary)
- Recommended file format : https://data.4tu.nl/info/fileadmin/user_upload/Documenten/Preferred_File_Formats_2019.pdf
  - The selected file format is recommended for data sharing, reuse and preservation.

What to check before publishing data/code?

Data documentation

- 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!

Example



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.

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

Methodology information, e.g. test equipment

Other information, e.g. organisation and naming convention

Yao & Alderliesten (2015). https://doi.org/10.4121/uuid:6da548f6-f801-41b4-8d88-db9ae81f6913

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

Examples

4TU.ResearchData

"Qualitative coding of 12 semi-structure interviews on food behaviour context and food reporting engagement": https://doi.org/10.4121/uuid:02b93c7c-545d-4501-b375-6db1aff039c6 (IDE)

"Transport Patterns of Global Aviation NOx and their Short-term O3 Radiative Forcing – A Machine Learning Approach". https://doi.org/10.4121/16886977.v1 (AE)

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 is published, and the means of publication are deemed reasonably by the supervisory team, it is fine
  - As long as the quality of the data is "sufficient" for the supervisory team, it is fine
- Not happy with the published datasets ?
  - In the 4TU.ResearchData repository you can create a new version of the dataset (same DOI!)

Available support

- Faculty Data Stewards
- 4TU.ResearchData 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 (personal data?): privacy-tud@tudelft.nl
  - Anything else? Not sure who to contact ? Check with your local data steward.

Q&A

CONTACT INFO:

Feedback and suggestions

Survey https://evasys-survey.tudelft.nl/evasys/online.php?p=967NX

Short URL https://tinyurl.com/36r43ehc

QR Code: