Schedule

- Introduction to Data Stewardship
- Break
- FAIR data
- Data Management Planning
- Break
- Writing a Data Management Plan session
- Final questions & remarks

TU Delft

Data Stewardship and Research Data Management

Dr.Santosh Ilamparuthi
S.Ilamparuthi@tudelft.nl

What is Data Stewardship?

Lots of Definitions…

The Data Management Body of Knowledge (DMBOK) defines Data Stewardship as:
“The most common label to describe accountability and responsibility for data and processes that ensure effective control and use of data assets. Stewardship can be formalized through job titles and descriptions, or it can be a less formal function driven by people trying to help an organization get value from its data.”

The Data Governance Institute:
“Data Stewardship is concerned with taking care of data assets that do not belong to the stewards themselves. Data Stewards represent the concerns of others. Some may represent the needs of the entire organization. Others may be tasked with representing a smaller constituency: a business unit, department, or even a set of data themselves.”

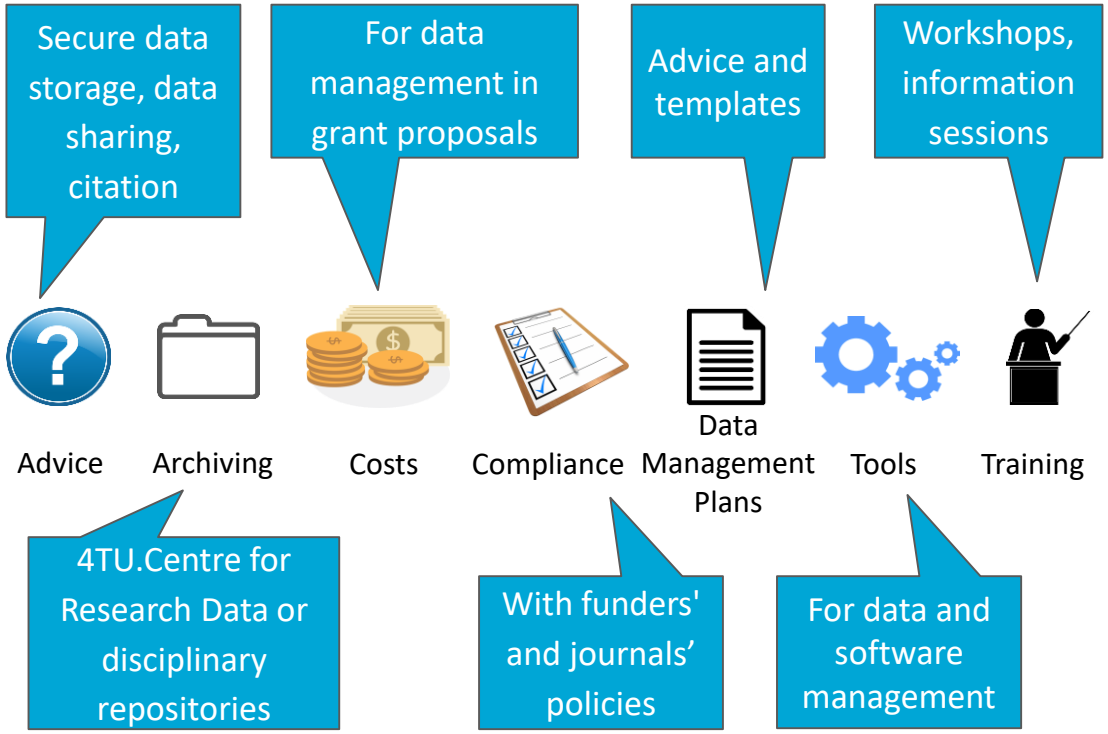
Lots of Definitions…

The Data Management Body of Knowledge (DMBOK) defines Data Stewardship as:
"The most common label to describe accountability and responsibility for data and processes that ensure effective control and use of data assets. Stewardship can be formalized through job titles and descriptions, or it can be a less formal function driven by people trying to help an organization get value from its data."

And many more!

The Data Governance Institute:
"Data Stewardship is concerned with taking care of data assets that do not belong to the stewards themselves. Data Stewards represent the concerns of others. Some may represent the needs of the entire organization. Others may be tasked with representing a smaller constituency: a business unit, department, or even a set of data themselves."

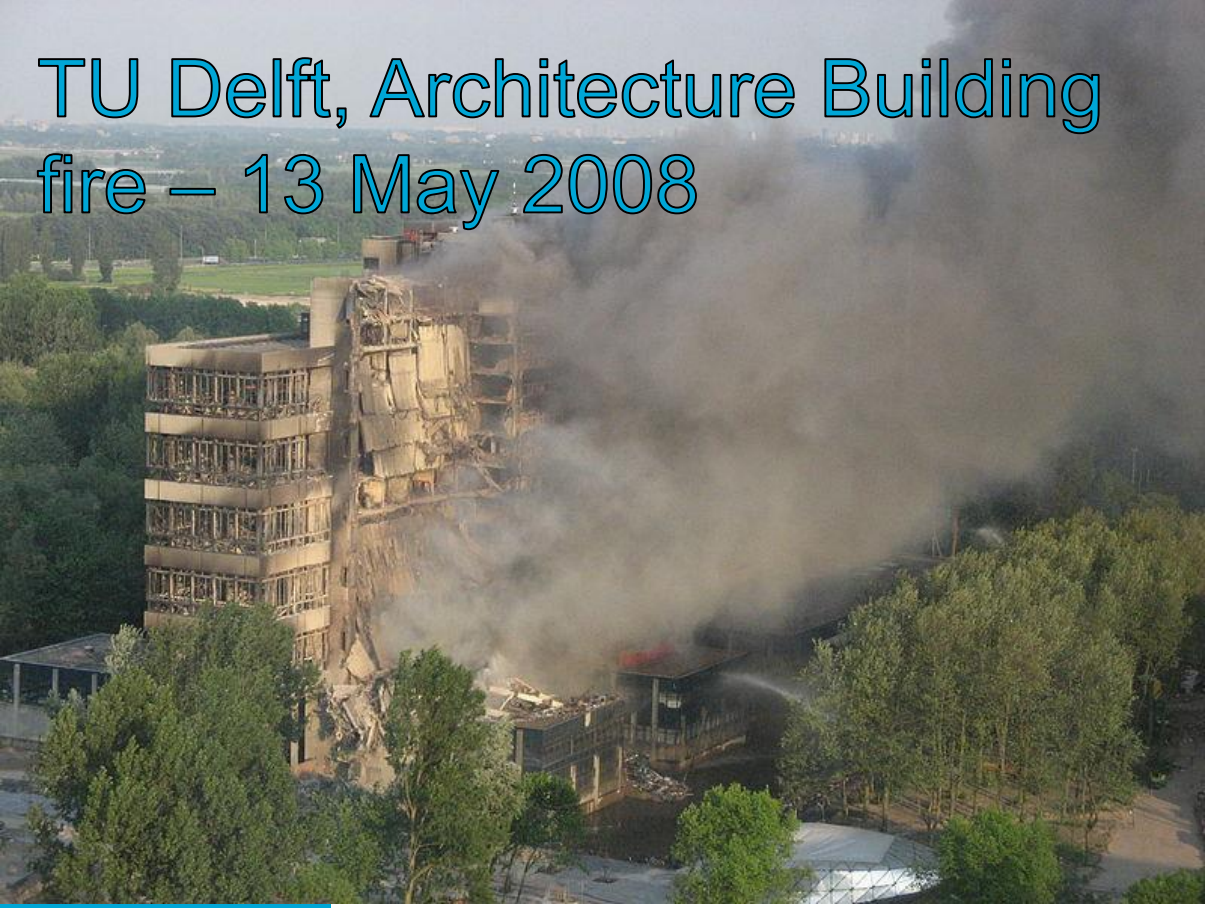
Data Stewardship at TU Delft



https://www.tudelft.nl/en/library/current-topics/research-data-management/r/support/data-stewardship/

What can Data Stewardship do for you – the Researcher?

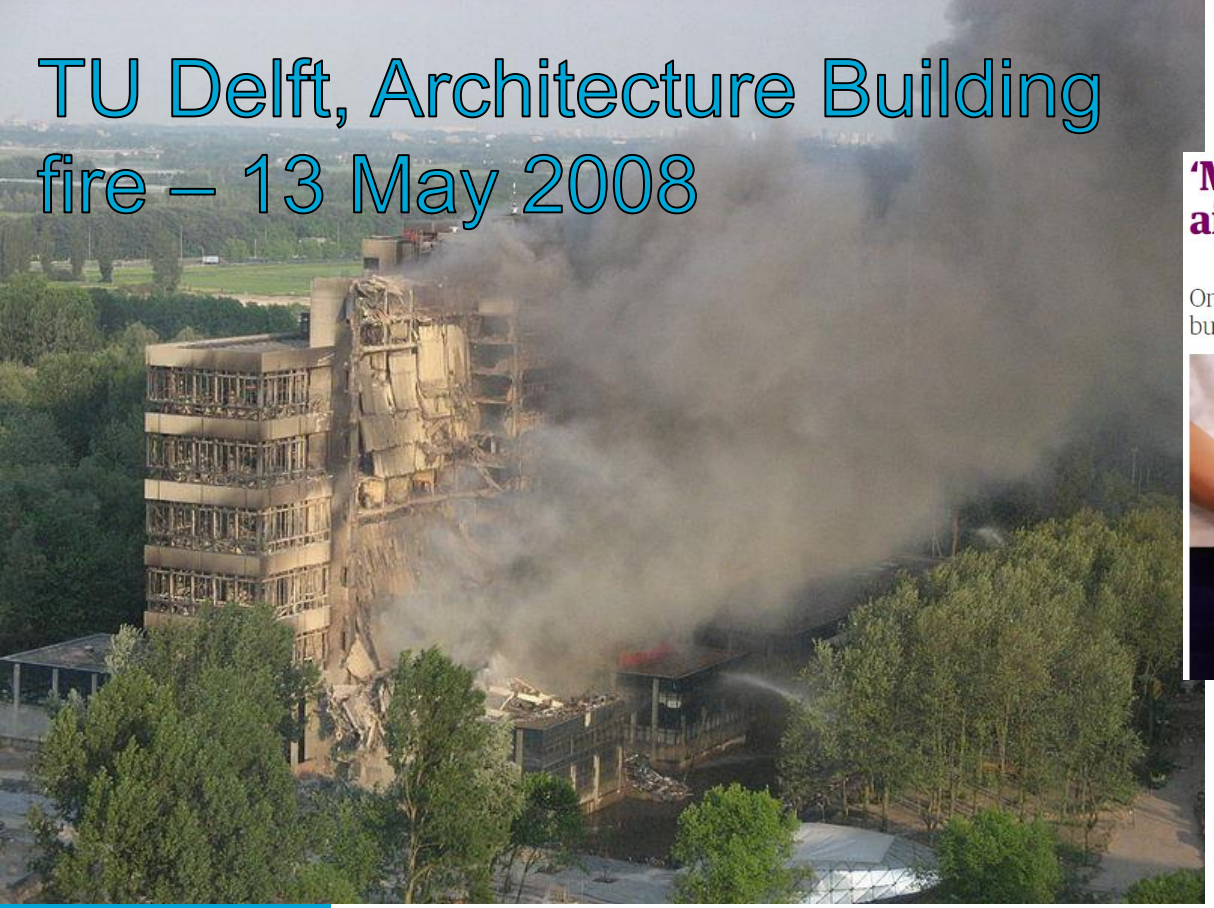
TU Delft, Architecture Building fire – 13 May 2008



https://commons.wikimedia.org/wiki/File:Brand_bouwkunde_-_TU_Delft_-_13_Mei_2008.jpg

TU Delft

TU Delft, Architecture Building fire – 13 May 2008



https://commons.wikimedia.org/wiki/File:Brand_bouwkunde_-_TU_Delft_-_13_Mei_2008.jpg

‘My £1,000 Macbook Air was stolen at airport security and no one cares’

One traveller found that some airports can identify thieves - but do nothing to chase them or return your goods

https://www.theguardian.com/money/2018/may/04/my-1000-macbook-air-was-stolen-at-airport-security-and-no-one-cares

How much of your data would you lose if… and how would you react?

- your laptop got stolen
- your lab burnt
- you lost your USB stick
- your portable hard drive got damaged
- Your stuff on Dropbox / Googledrive disappeared

NOOOO S”*f*!!!**

It hurts, But I can manage with some work

Easy peasy, I have it covered

Original Slide by Marta Teperek

How much of your data would you lose if… and how would you react?

- your laptop got stolen
- your lab burnt
- you lost your USB stick
- your [illegible]
- You [illegible] disappeared

I have earned my stripes the hard way – My horror story ☹

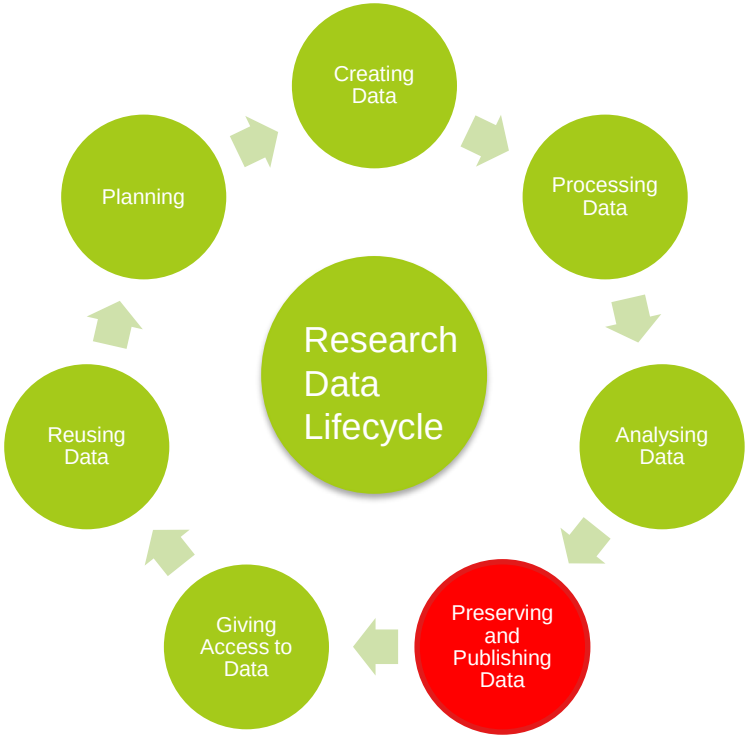
NOOOO S”*f*!!!**

It hurts, But I can manage with some work

Easy peasy, I have it covered

Original Slide by Marta Teperek

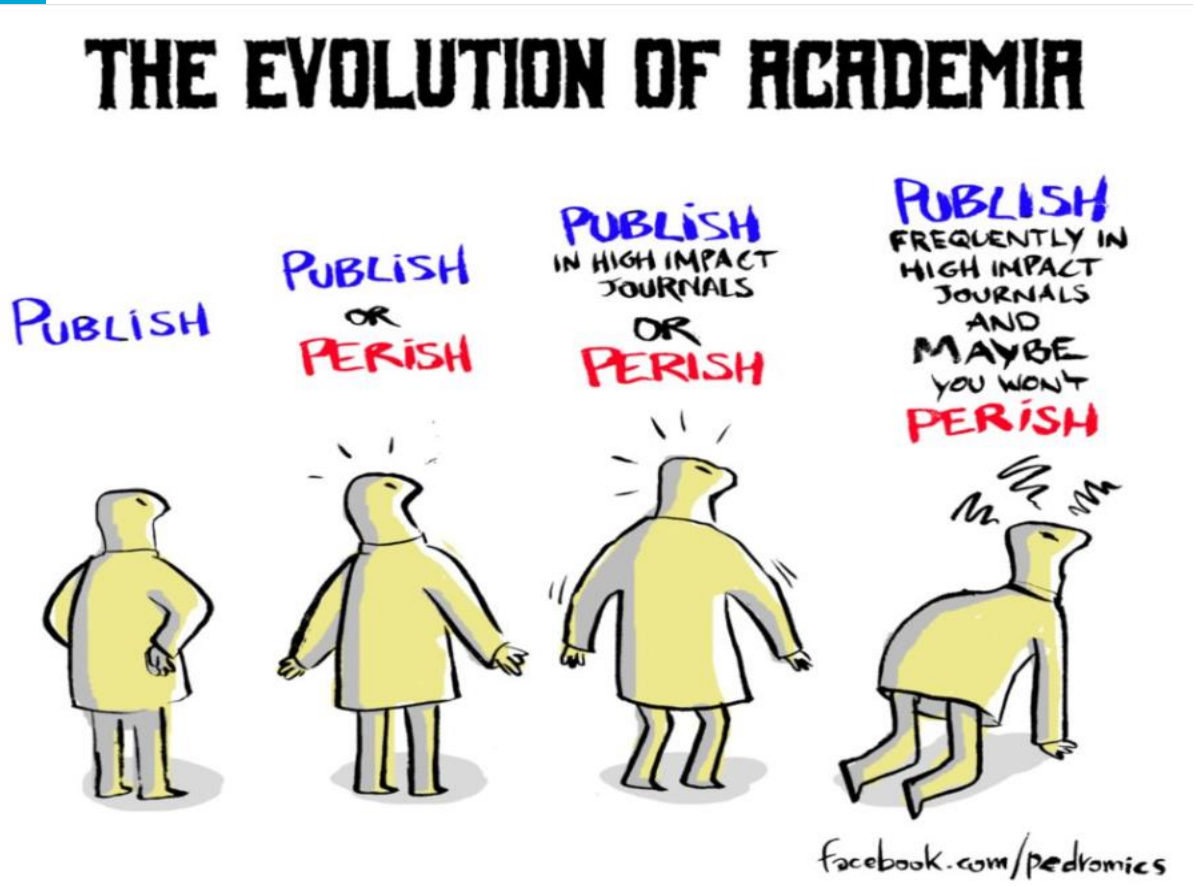
Research Data Lifecycle



The only thing that counts in academia is publication of novel results in high impact journals

https://www.repository.cam.ac.uk/handle/1810/276106

Based on UK data archive lifecycle and slide by Marta Teperek



Copyright: https://velica.deviantart.com/art/Publish-or-Perish-645355248

Fraud!!

Report finds massive fraud at Dutch universities

Investigation claims dozens of social-psychology papers contain faked data.

Ewen Callaway

When colleagues called the work of Dutch psychologist Diederik Stapel too good to be true, they meant it as a compliment. But a preliminary investigative report (go.nature.com/tqmp5c) released on 31 October gives literal meaning to the phrase, detailing years of data manipulation and blatant fabrication by the prominent Tilburg University researcher.

Fraud!!

Report finds massive fraud at Dutch universities

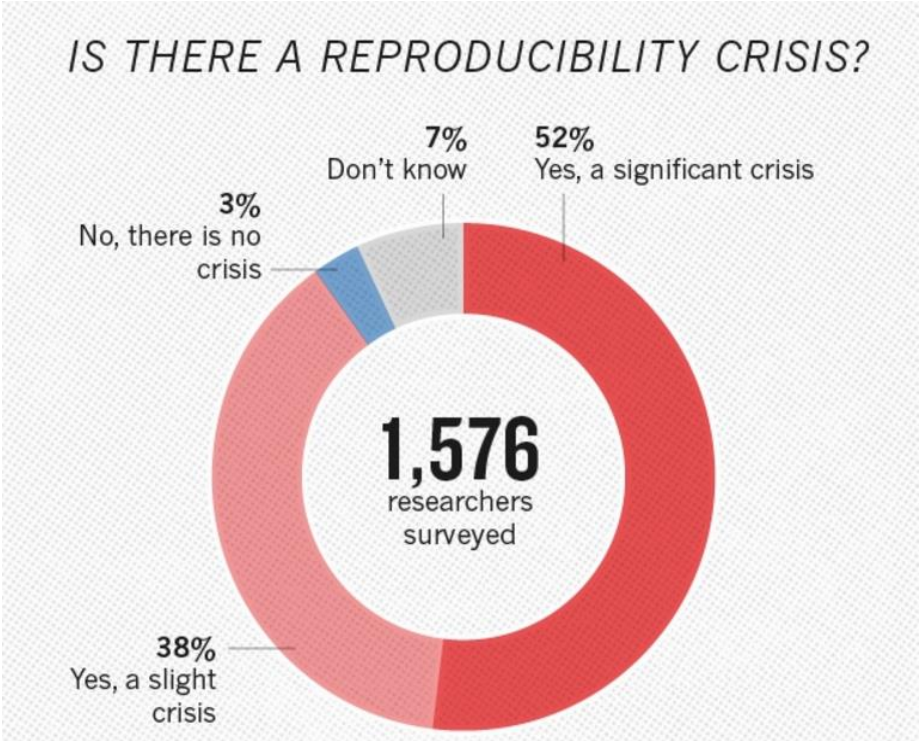
Investigation claims dozens of … contain faked data.

Schön Scandal!!

Ewen Callaway

When colleagues called the work of Dutch psychologist Diederik Stapel too good to be true, they meant it as a compliment. But a preliminary investigative report (go.nature.com/tqmp5c) released on 31 October gives literal meaning to the phrase, detailing years of data manipulation and blatant fabrication by the prominent Tilburg University researcher.

Reproducibility Crisis



https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

The main reasons for this crisis are

- Selective reporting
- Pressure to publish
- Insufficient supervision and training
- **Supporting data / methods / code not available**

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature, [online] 533(7604), pp.452-454. Available at: https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970 [Accessed 26 Apr. 2018].

Funder	BMBF	DFG	EU Horizon 2020	Wellcome Trust
What to archive?	research data	research data, research software, supplemental material	reserach data, unpublished data, programming code	reserach data
Where to archive?	data archive/repository	free choice of repository	free choice of repository	no specific stipulation, but must be archived
When to archive?	after project end	within 12 months after project ends	as soon as possible	immediately after publication
General terms and conditions	data need to be longterm-archived/ preserved for the scientific community	primary data have to be stored for 10 years in their institution of origin	a data management plan and making data openly available after the project ends is mandatory for all projects with start date after 01/2017	mandatory for all projects whose data can be shared and re-used; re-users have to cite the data source and keep to the re-use license conditions

Table and information, E.Böker “https://www.forschungsdaten.info/praxis-kompakt/english-pages/funder-guidelines/”

Funder	BMBF	DFG	EU Horizon 2020	Wellcome Trust
				eserach data
				no specific stipulation, but must be archived
When to archive?	after project end	within 12 months after project ends	as soon as possible	immediately after publication
General terms and conditions	data need to be longterm-archived/ preserved for the scientific community	primary data have to be stored for 10 years in their institution of origin	a data management plan and making data openly available after the project ends is mandatory for all projects with start date after 01/2017	mandatory for all projects whose data can be shared and re-used; re-users have to cite the data source and keep to the re-use license conditions

DFG and the European Commission require in their guidelines that applicants submit a “data management plan” (DMP) along with their application which describes how research data will be managed and shared during and after the research project

Table and information, E.Böker “https://www.forschungsdaten.info/praxis-kompakt/english-pages/funder-guidelines/”

TU Delft

Funder	BMBF	DFG	EU Horizon 2020	Wellcome Trust
				eserach data
				no specific stipulation, but must be archived
When to archive?	after project end	within 12	as soon as possible	immediately after pu-
General terms and conditions	data need to be longterm-archived/ preserved fo the scientific community		institution of origin	mandatory for all projects with start date after 01/2017	to cite the data source and keep to the re-use license conditions

DFG and the European Commission require in their guidelines that applicants submit a “data management plan” (DMP) along with their application which describes how research data will be managed and shared during and after the research project

The DFG “Guidelines on Safeguarding Good Scientific Practice (2019)” specify that data have to be managed according to the F.A.I.R. principles.

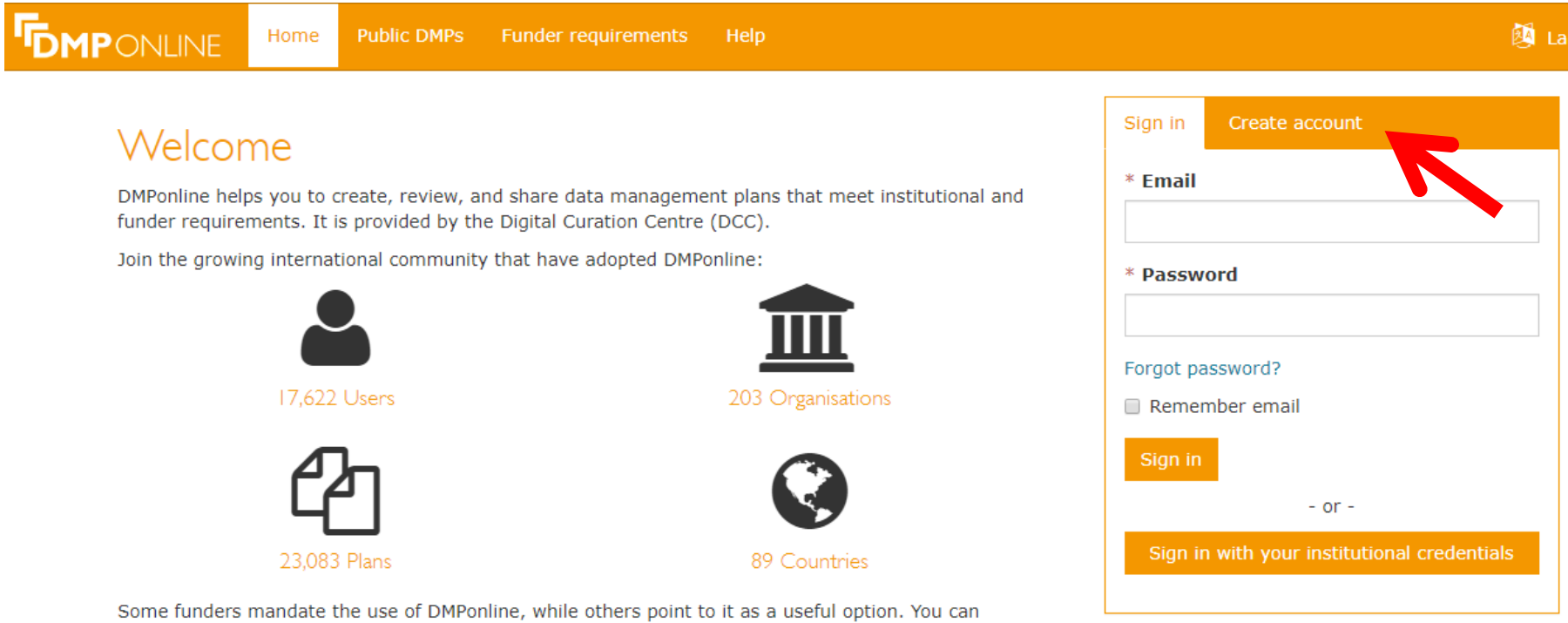
Table and information, E.Böker “https://www.forschungsdaten.info/praxis-kompakt/english-pages/funder-guidelines/”

TU Delft

Break time!

DMPOnline

https://dmponline.dcc.ac.uk/



TU Delft

FAIR principles

Esther Plomp
E.Plomp@tudelft.nl
@PhDToothFAIRy

https://the-turing-way.netlify.app/welcome

Image
https://zenodo.org/record/3695300#.XvoPYnJS-Uk

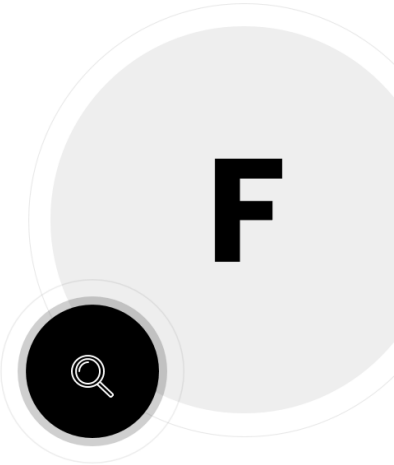
Original slides:
https://doi.org/10.5281/zenodo.3471553

The FAIR principles are a set of guidelines to increase the re-use of your data

Findable

- Deposit your data in a data repository with metadata and a persistent identifier

Findable



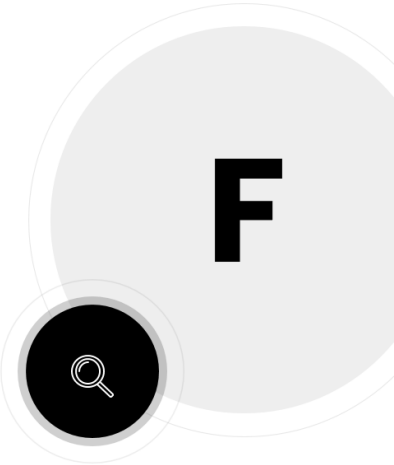
- Deposit your data in a **data repository** with metadata and a persistent identifier

an online archive that curates research datasets and provides long-term access

- Finalised datasets
- ~10-15 years

DRYAD zenodo figshare

Findable



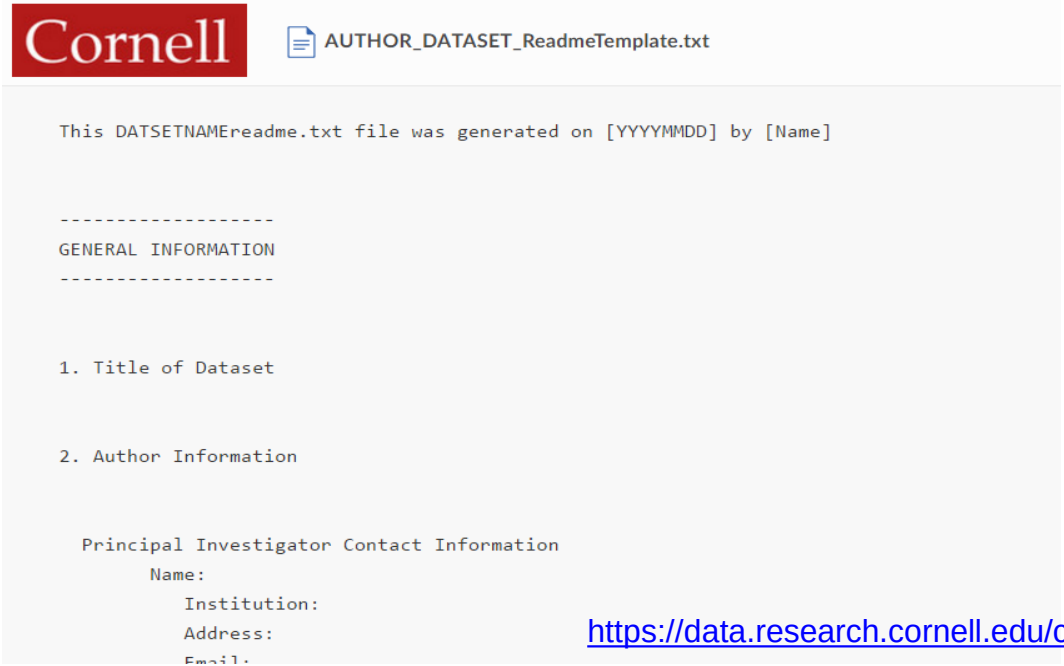
- Deposit your data in a data repository with **metadata** and a persistent identifier

Metadata = Information about data

- Contextual information
  - Title, author, keywords
  - When? For what purpose?
  - Size? Standards?

Metadata

Use a README file to put your data into context:



https://data.research.cornell.edu/content/readme

README files template: https://cornell.app.box.com/v/ReadmeTemplate

Evaluation of neodymium isotope analysis of human dental enamel as a provenance indicator using 10^{13} Ω amplifiers (TIMS)

E. Plomp [a], I.C.C. von Holstein [a], J.M. Koornneef [a], R.J. Smeets [a], J.A. Baart [b, c, 1], T. Forouzanfar [b, c], G.R. Davies [a]

Show more

https://doi.org/10.1016/j.scijus.2019.02.001

Get rights and content

Under a Creative Commons license

open access

https://doi.org/10.1016/j.scijus.2019.02.001

- Deposit your data in a data repository with metadata and a **persistent identifier**

ORCID

A **persistent identifier** (PI or PID) is a long-lasting reference to a file, web page, or other object

TU Delft

Accessible

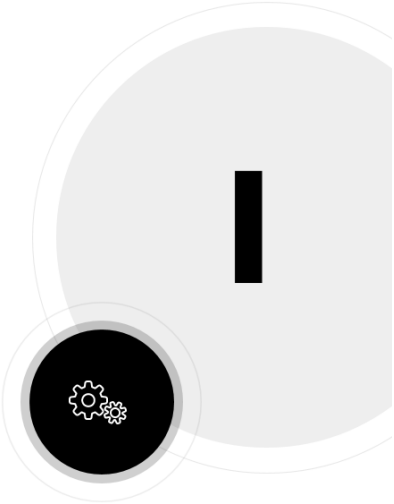
A

- Consider what will be shared (metadata)
- Obtain participant consent and perform risk management
- Determine access control

Interoperable

- Use open/common format/languages
- Consistent vocabulary
- Discipline common metadata standards
  - FAIRsharing.org
  - Research Data Alliance metadata directory
  - Digital Curation Center

Interoperable



Standard Data Formats

“widely accepted”
e.g., .xls .xlsx

Open Data Formats

Free to use
e.g., .csv .tab

Using standard or open data formats ensures longer-term usability of data

Convert files into standard open formats for long term preservation of the data

Original slide by Heather Andrews Mancilla

TU Delft

Data Formats

Preferred (sustainable) file formats that receive full preservation	
Text	Plain text, XML, HTML, PDF (PDF/A-1), JSON, PDB (Protein Data Bank), XYZ *(all formats should be encoded in UTF-8)*
Spreadsheets	CSV (Comma-separated values), Tab-delimited values, PDF (PDF/A-1)
Images	JPEG, TIFF, PNG, SVG
Geospatial	GML (Geographical Mark-up Language), KML (Keyhole Mark-up Language), ESRI Shapefile, Geo-referenced TIFF
Numerical	NetCDF, CSV, JSON
Video	*No sustainable format established*
Audio	Waveform Audio File Format (WAVE)
Databases	Delimited Flat File w/DDL
Archives	ZIP, TAR, GZIP, 7Z

https://researchdata.4tu.nl/fileadmin/user_upload/Documenten/preffered_file_formats.pdf

Slide adapted from Christine Staiger https://doi.org/10.5281/zenodo.2585691

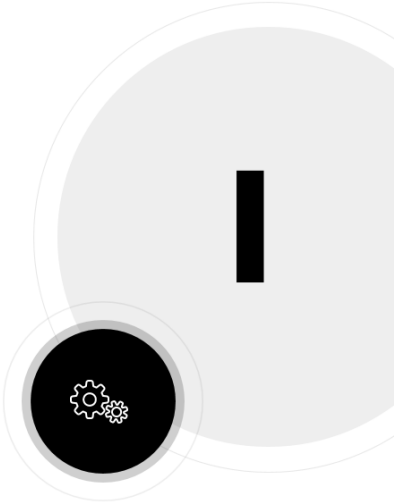
Metadata standard: how to say 'female'

18-day pregnant females
2 yr old female
400 yr. Old female
Adult female
Asexual female
Castrate female
Cf.female
Cystocarpic female
Dikaryon
Dioecious female
Diploid female
F
Famale
Femail
Female
Female – worker
Female (alate sexual)
Female (calf)
Hen
Female (lactating)
Female (pregnant)
Female (outbred
Female parent
Female plant
Female with eggs
Female worker
Female, 6-8 weeks old
Female, virgin
Female, worker
Female(gynoecious)
Femele
Female, pooled
Femalen
Females
Females only
Gynoecious
Healthy female
Probably female (morphology)
Individual female
Igb*cc females
Mare
Female (worker)
Monosex female
Ovigerous female
Oviparous sexual females
Worker bee
Female enriched
Pseudohermaprhoditic
female
Remale
Semi-engorged female
Sexual oviparous female
Sterile female worker
Strictly female
Tetraploid female
Thelytoky
Female (gynoecious)
Worker caste 'female'
Sex female
Female, other
Female child
Femal
3 female
Female (phenotype)
Female mice
Female, spayed
Femlale
Metafemale
Sterile female
Normal female
Sf
Vitellogic replate female
Worker
Hexaploid female
Female (f-o)

Metadata Standard

Dublin Core Metadata:

Creator (Author): a person, an organization, or a service
Contributor: a person, an organization, or a service
Title: name given to the resource
Date: e.g., YYYY-MM-DD
Language: use a controlled vocabulary such as RFC 4646, e.g. nl-NL
Format: file format, physical medium, or dimensions of the resource
Subject: use keywords, key phrases, or classification codes
Description: may include an abstract, a table of contents, etc.
Identifier/Relation: formal identification of the resource
Source: a resource from which the described resource is derived
Rights: statement about various property rights / license
Publisher: include a person, an organization, or a service
Type: nature or genre of the resource
Coverage: spatial (e.g., geographic coordinates) or temporal (e.g., a named period, date, or date range) topic of the resource



Original slide by Heather Andrews Mancilla

http://dublincore.org/documents/2012/06/14/dcmi-terms/

Reusable

- Apply a licence to specify how others can reuse your data
- Documentation

Reusable

R

- Apply a **licence** how others can your data
- Documentation

Licences for data

- Public Domain Dedication (CC0)
- Attribution (CC BY)
- Attribution-NoDerivatives (CC BY-ND)
- Attribution-NonCommercial (CC BY-NC)
- Attribution-NonCommercial-ShareAlike (CC BY-NC-SA)
- Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND)

Licences for software and code

- MIT License
- Apache Licence 2
- GNU General Public Licence 3 (GNU GPLv3)

https://creativecommons.org/licenses/

https://researchdata.4tu.nl/en/use-4turesearchdata/archive-research-data/upload-your-data-in-our-data-archive/licencing/

TU Delft

Data Repositories & FAIR

Finalised Datasets
Snapshot

Long-term preservation
~10-15 years

Findable
- DOI
- Metadata

Accessible
- Control

Interoperable
- Metadata
- Vocabulary
- Open formats/standards

Reusable
- Licence

Repository for code/software

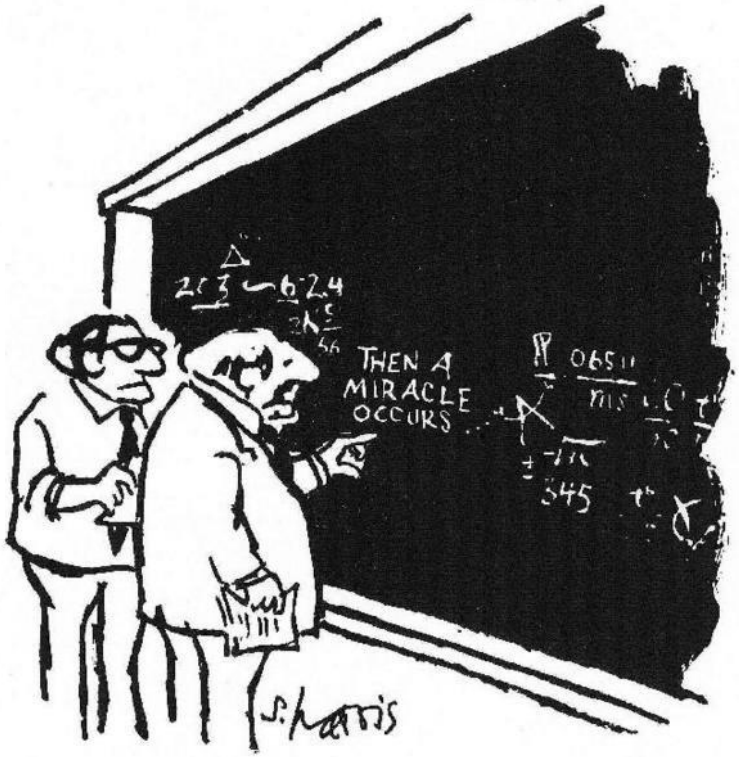
Software:

https://www.youtube.com/watch?v=pjsbBQYOOaE&t=1s

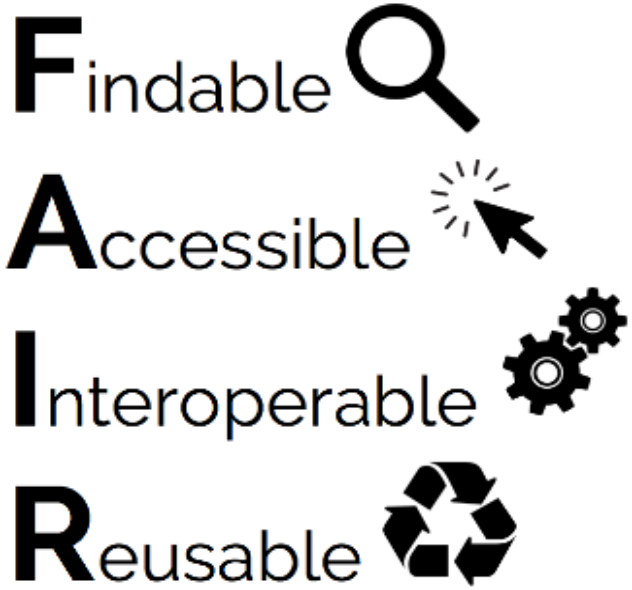
https://guides.github.com/activities/citable-code/

Repository for protocols

protocols.io



"I think you should be more explicit here in step two."



≠

Standard

Open

Intrinsic quality

More info?

- Go FAIR
- Top 10 FAIR Data/Software Things
- F1000 FAIR Guide
- The FAIR Guiding Principles for scientific data management and stewardship (2016)

https://doi.org/10.5281/zenodo.2667392

Data Management Planning

Dr. Heather Andrews
H.E.AndrewsMancilla@tudelft.nl

What is a Data Management Plan?

Image by Peggy und Marco Lachmann-Anke from Pixabay

Data Management Plan

- The Data Management Plan (DMP) is a document where you state how you will **manage the research data/code during research** and **after the research** project is finished.

- A DMP is a **living document**.

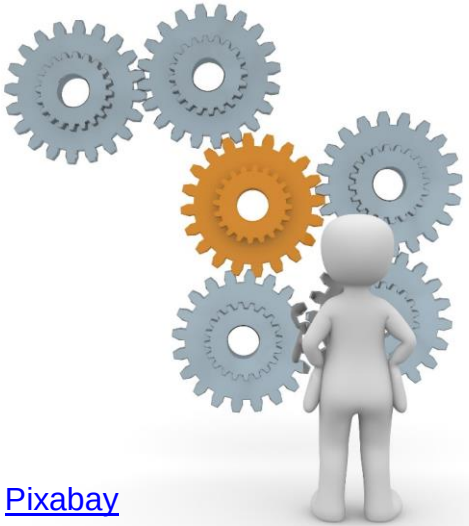


Image by Peggy und Marco Lachmann-Anke from Pixabay

Why a Data Management Plan?

Image by Peggy und Marco Lachmann-Anke from Pixabay

Awareness!

Results TU Delft Research Data Management Survey
for the Faculty of Aerospace engineering
Results 2017 published in https://doi.org/10.18352/lq.10287
Results 2019 in prep.

It helps you to **organize your work**
&
It also helps us (support) to have a more **concrete view** of what is missing!

Respondents that work with DMPs appear to be more aware of FAIR principles and more aware (and use!) data repositories

Respondents working in projects **WITH DMPs**

Year	Use repos	Use or are aware of repos	Data loss	Aware of FAIR
2017	27%	81%	15%	50%
2019	**50%**	**92%**	**8%**	**67%**

Respondents working in projects **WITHOUT DMPs**

Year	Use repos	Use or are aware of repos	Data loss	Aware of FAIR
2017	16%	63%	13%	17%
2019	**24%**	**66%**	**12%**	**38%**

TU Delft

How to prepare a DMP?

Image by Peggy und Marco Lachmann-Anke from Pixabay

DMP Preparation – During Research

- **Who is the owner of the data?**
- **Which policies apply?**
- **What type of data/code will be collected and/or generated?**

TU Delft

Data Types

Experiments
Simulations-modelling data
Software (scripts, self-standing)
Interviews, recordings, surveys
Human participants data
Design data (maps, charts)
Etc.

Standard

Sensitive

Personal
Commercial
Dangerous

DMP Preparation – During Research

- Who is the owner of the data?
- Which policies apply?
- What type of data/code will be collected and/or generated?
- **Where will the data be stored?**
- **What will be the backup strategy?**
- **How will you transfer data/code with collaborators?**
- **What software, equipment, facilities do you need?**

TU Delft

DMP Preparation – During Research

- Who is the owner of the data?
- Which policies apply?
- What type of data/code will be collected and/or generated?
- Where will the data be stored?
- What will be the backup strategy?
- How will you transfer data/code with collaborators?
- What software, equipment, facilities do you need?
- **How will you organize the data/code? (e.g., directory structure, file-naming conventions, version control)**
- **How will you document the data/code?**

TU Delft

DMP Preparation – After Research

- **Which data/code will be archived for the long-term?**
  - All research output necessary to **validate results presented in a journal article** and all research output **suitable for reuse**.

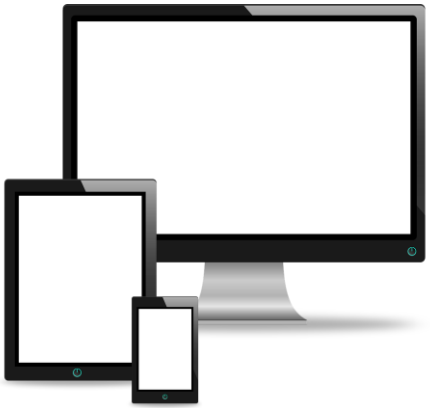


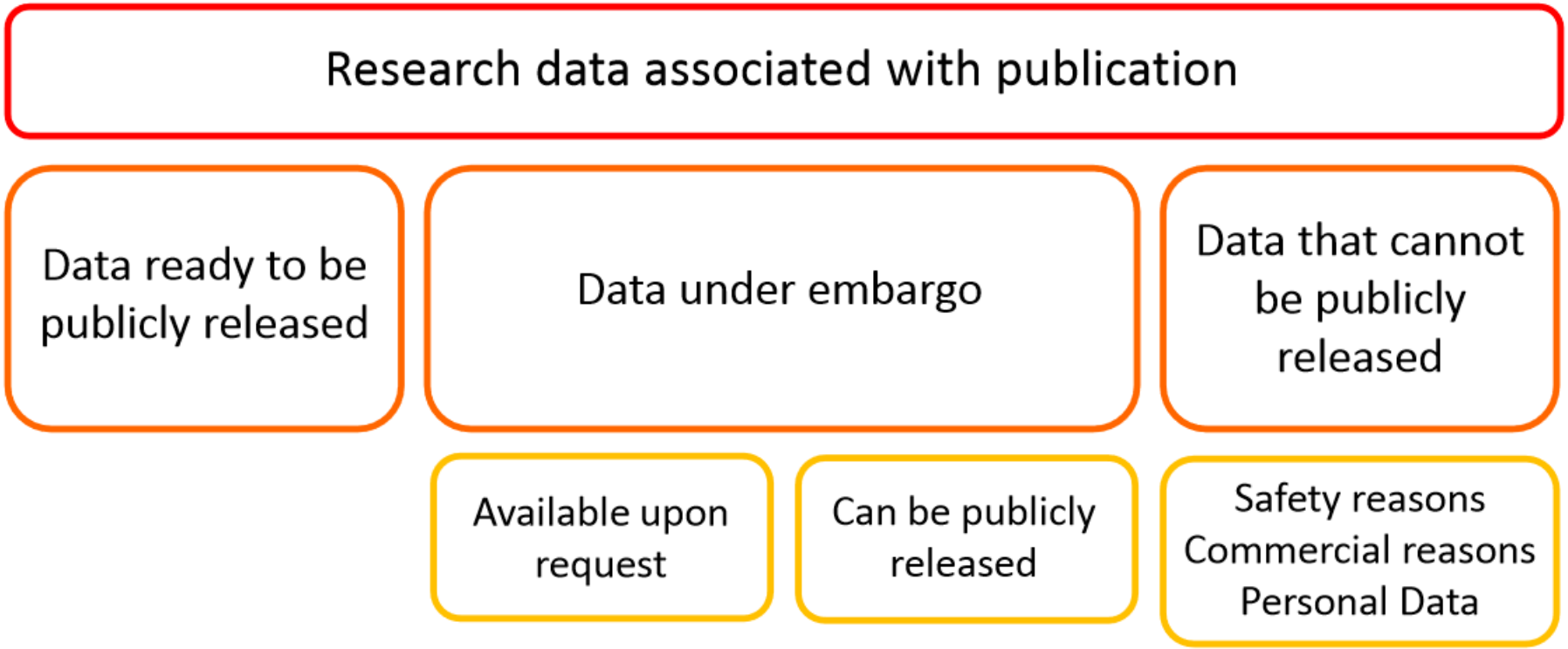
Image by mmi9 from Pixabay

Image by OpenClipart-Vectors from Pixabay

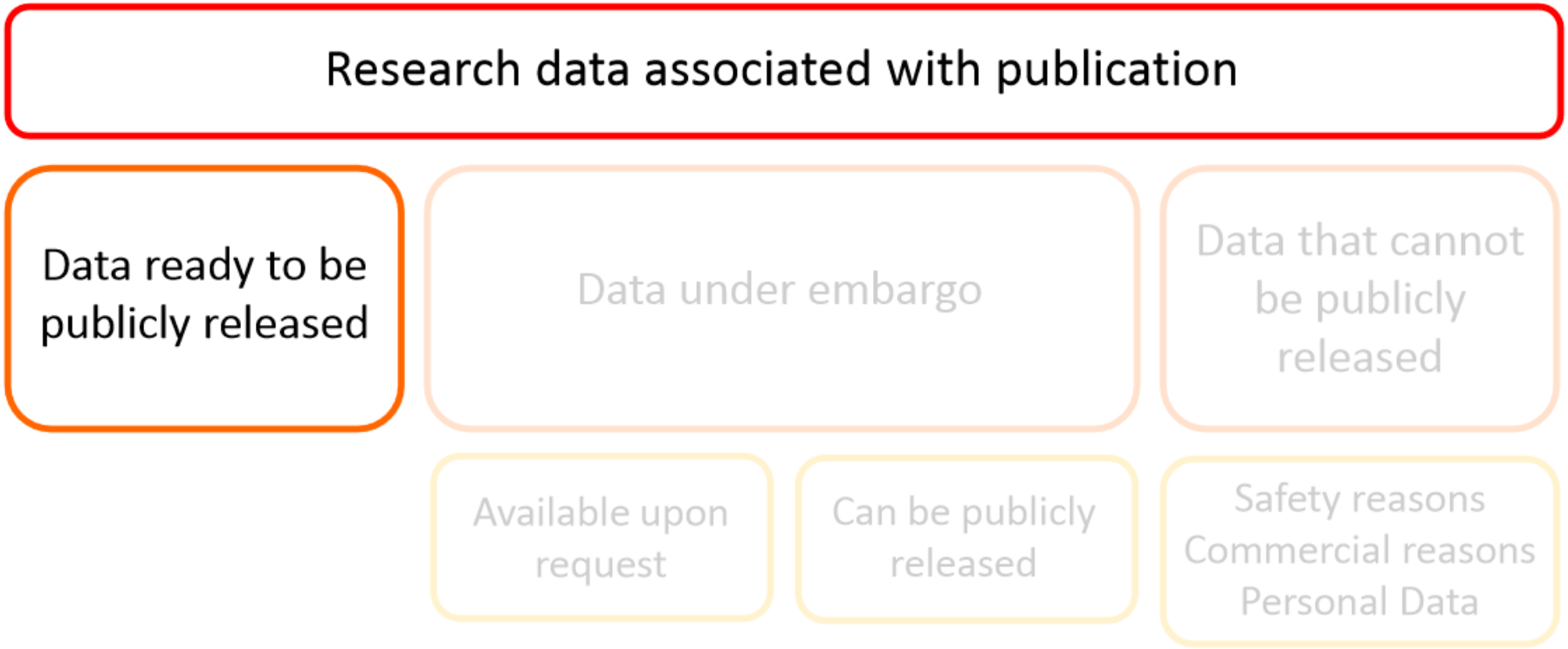
DMP Preparation – After Research

- **Which data/code will be archived for the long-term?**
  - All research output necessary to **validate results presented in a journal article** and all research output **suitable for reuse**.
- **How will the data/code be archived for the long-term?**

TU Delft



Here ‘data’ includes code



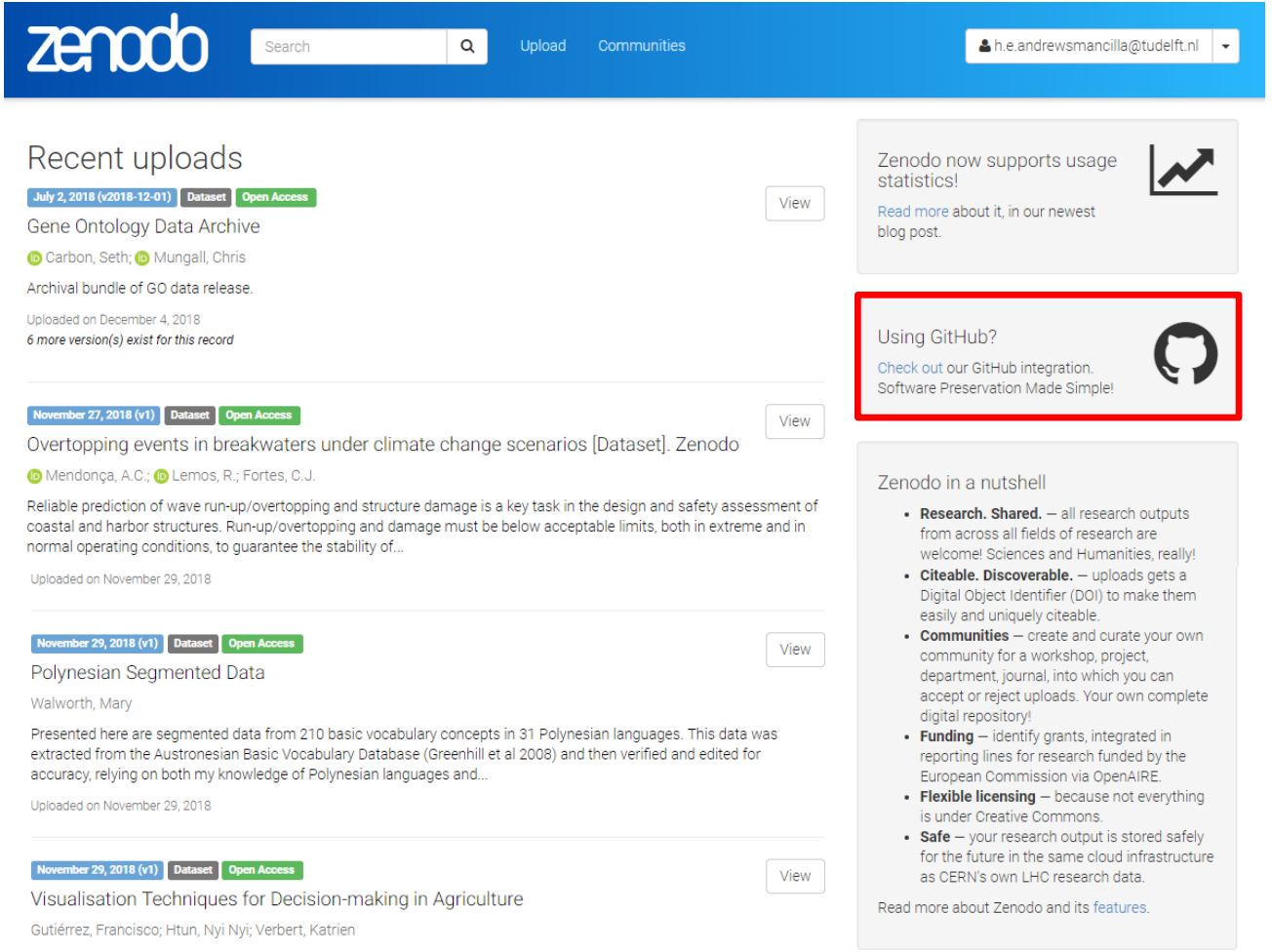
Here ‘data’ includes code

Zenodo

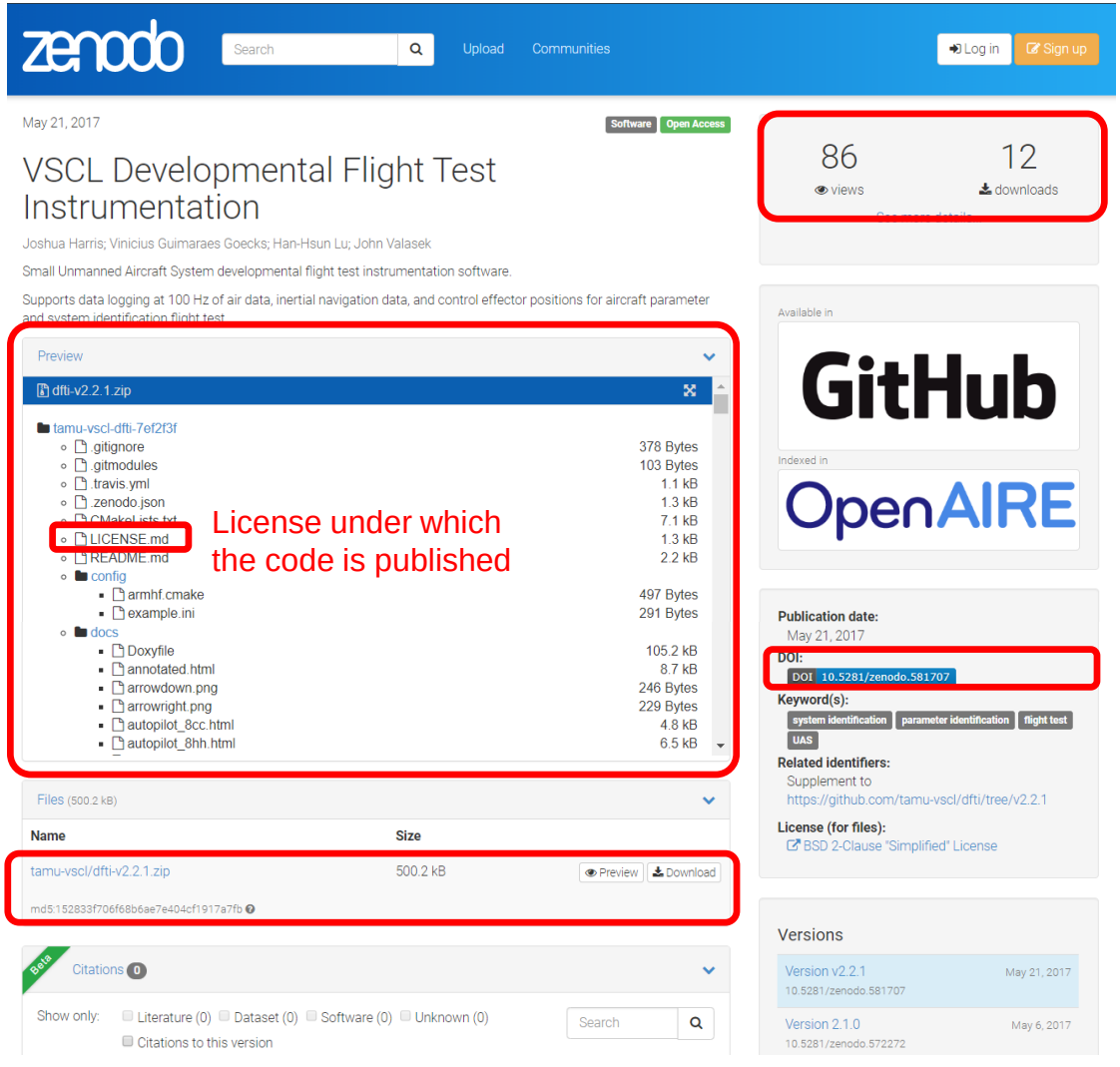
https://zenodo.org/

Suitable for data and code

Github integration
→ Zenodo archives a snapshot of the Github repository



TU Delft



Preview of content
In this case, a snapshot of a Github repository

License under which the code is published

Easy download

How many downloads

Persistent identifier (code becomes citable)

TU Delft

DMP Preparation – After Research

- Which data/code will be archived for the long-term?
  - All research output necessary to validate results presented in a journal article and all research output suitable for reuse.
- How will the data/code be archived for the long-term?
- **Who will be the contact person?**

TU Delft

How to start writing a DMP?

Image by Peggy und Marco Lachmann-Anke from Pixabay

Different funders/institutions
Different DMP templates!

Image by Andrew Martin from Pixabay

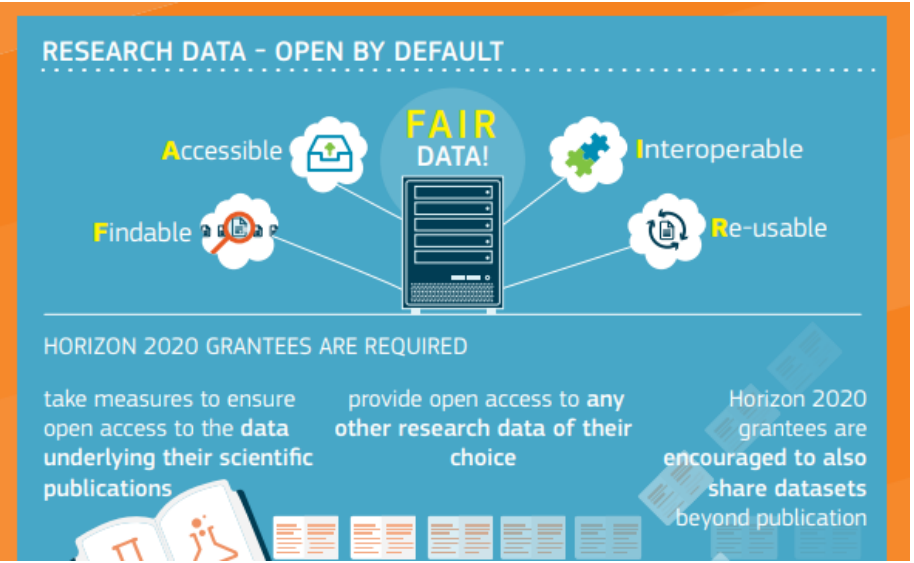
Horizon 2020 DMP

- Horizon 2020 (H2020) is the current and biggest EU Research and Innovation programme (2014 to 2020) focused on funding excellent science, industrial leadership and tackling societal challenges.

- Great focus on Open Science.

- Open Research Data Pilot
  - optional in H2020
  - mandatory in next framework Horizon Europe (2021-2027)

Horizon 2020



http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf#view=fit&pagemode=none

Horizon Europe

Open Science across the programme

Open Science — Better dissemination and exploitation of R&I results and support to active engagement of society

Mandatory Open Access to publications: beneficiaries shall ensure that they or the authors retain sufficient intellectual property rights to comply with open access requirements

Open Access to research data ensured: in line with the principle "as open as possible, as closed as necessary"; Mandatory Data Management Plan for FAIR (Findable, Accessible, Interoperable, Re-usable) and Open Research Data

https://ec.europa.eu/info/sites/info/files/research_and_innovation/strategy_on_research_and_innovation/presentations/horizon_europe_en_investing_to_shape_our_future.pdf

TU Delft

H2020 DMP

Based on **FAIR principles**

Initial DMP
Detailed DMP
Final Review DMP

Not only when **publishing data** but also **during the project!**

1. Data summary

2. FAIR data
    - 2.1 Making data findable, including provisions for metadata
    - 2.2 Making data openly accessible
    - 2.3 Making data interoperable
    - 2.4 Increase data re-use (through clarifying licenses)

3. Allocation of resources

4. Data security

5. Ethical aspects

6. Other

TU Delft

Next Step:
Let's start writing a DMP!

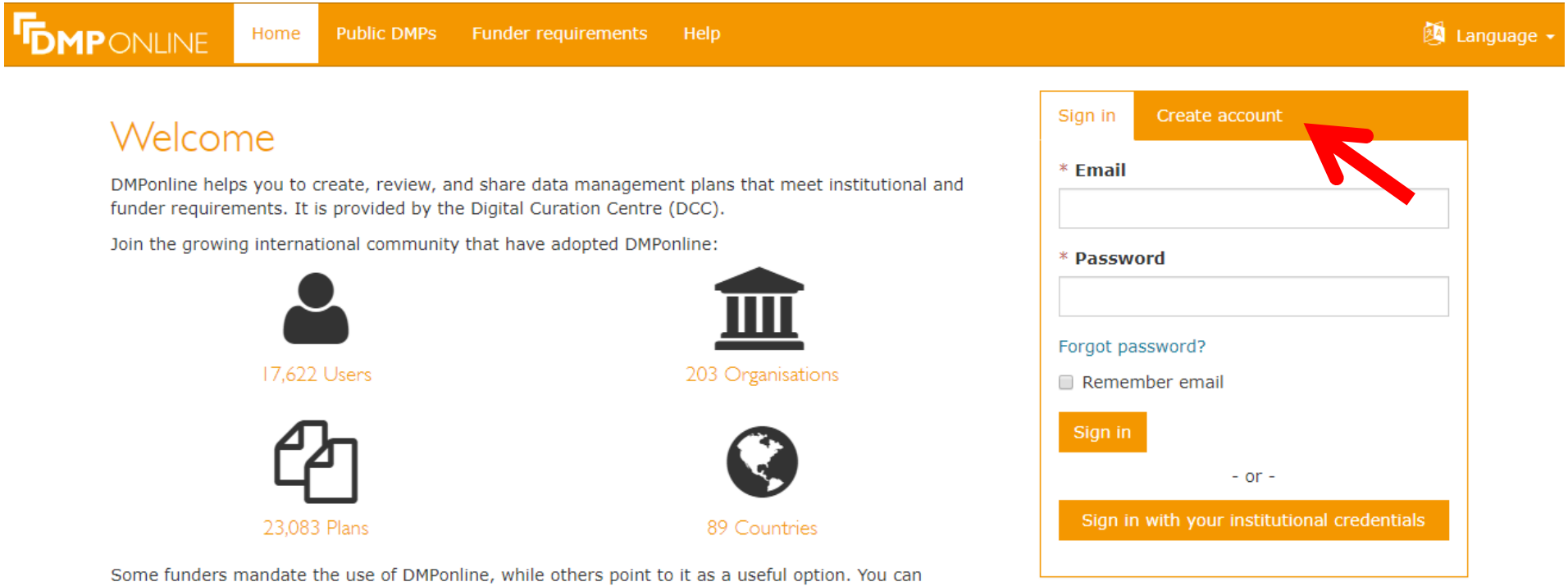
Image by Gordon Johnson from Pixabay

But after the break!

Image by Gordon Johnson from Pixabay

DMPOnline

https://dmponline.dcc.ac.uk/



TU Delft

Extra links

Data and Code licenses:

- https://creativecommons.org/licenses/
- https://choosealicense.com/
- https://www.gnu.org/licenses/license-recommendations.html

Archiving of data and/or code:

- https://softwaresaved.github.io/software-deposit-guidance/

Discipline specific metadata standards:

- http://rd-alliance.github.io/metadata-directory/subjects/engineering.html
- https://link.springer.com/article/10.1057%2Fdam.2010.29
- http://www.dcc.ac.uk/resources/metadata-standards

Repositories:

- https://www.re3data.org/
- https://guides.github.com/activities/citable-code/