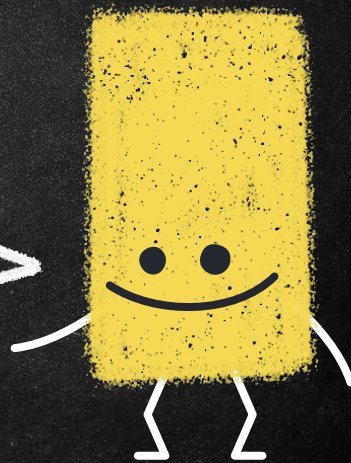
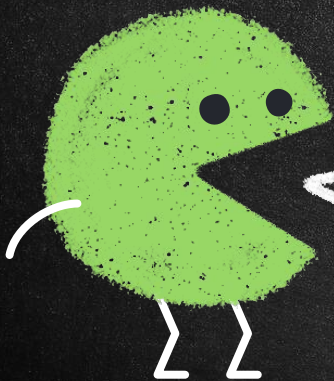




Esther Plomp
2021
[https://doi.org/
10.5281/zeno
do.4638744](https://doi.org/10.5281/zenodo.4638744)

DATA MANAGEMENT



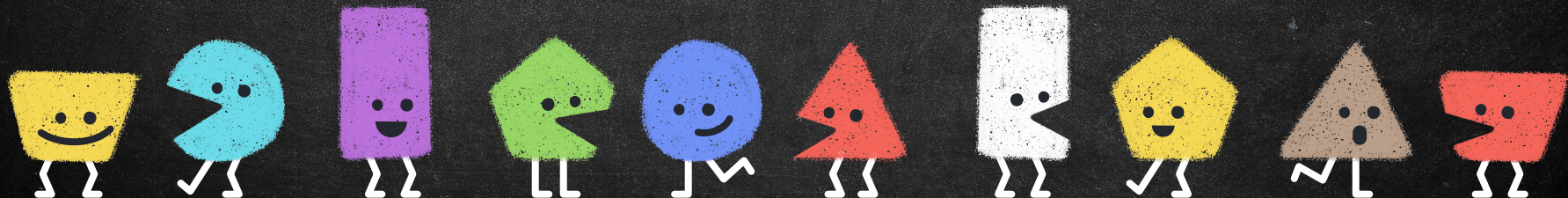


ESTHER PLOMP

Data Steward - Delft University of Technology

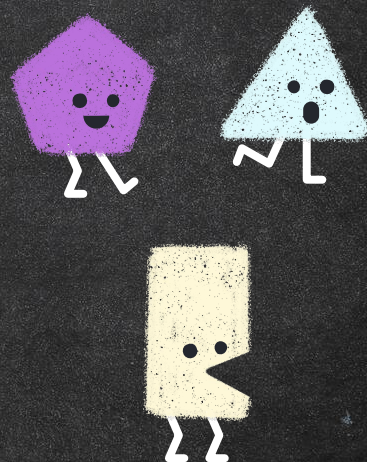
@PhDToothFAIRy - e.plomp@tudelft.nl

she/her

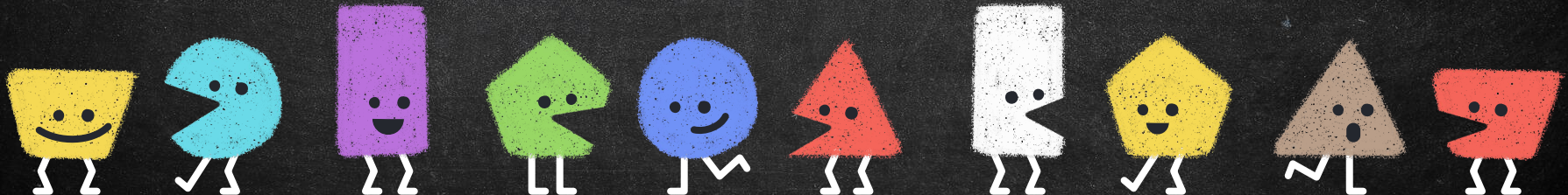


AGENDA

- Intro data management (15 min)
- Interactive notes (10 min)
- Break out session 1 (15 min)
- Break
- Break out session 2 (15 min)
- Intro open data (15 min)



INTRO DATA MANAGEMENT

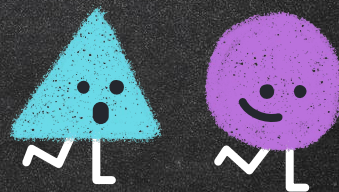
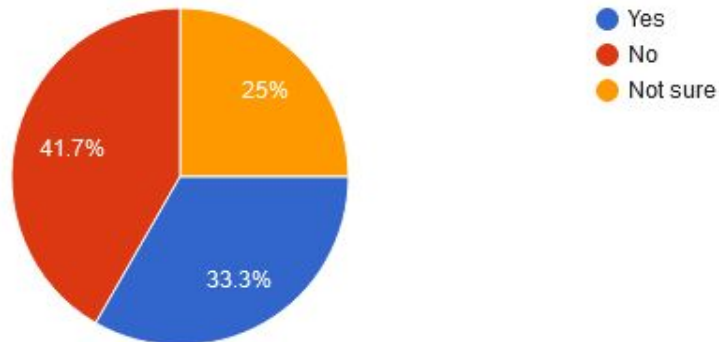


SLIDES ARE AVAILABLE

<https://doi.org/10.5281/zenodo.4638744>

Did you ever use a data repository?

12 responses

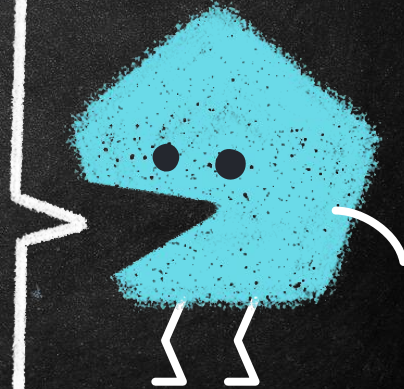
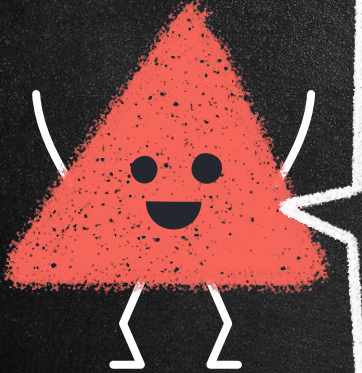


RESEARCH DATA

ANY TYPE OF INFORMATION THAT IS
COLLECTED, OBSERVED, OR CREATED, IN THE
CONTEXT OF RESEARCH

Raw/primary data: The data
measured/recorded without manipulation

Processed/secondary data: Data that has
been modified/processed for analysis



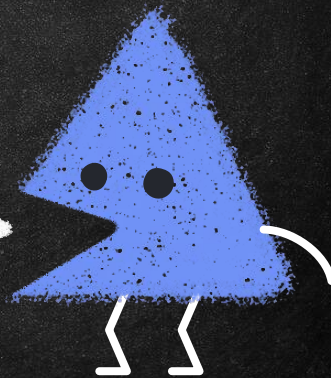
DATA MANAGEMENT PLAN

- Roles and responsibilities
- Type of data, documentation
- Storage, back-up solutions
- Preservation
- Reuse

The Turing Way -
Data Management
Plans

DMPonline

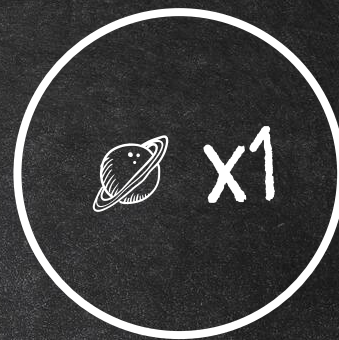
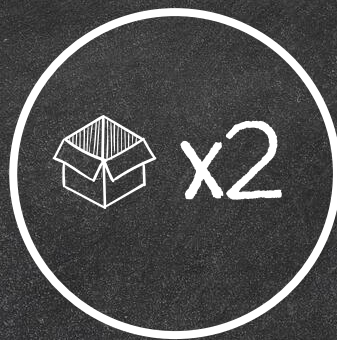
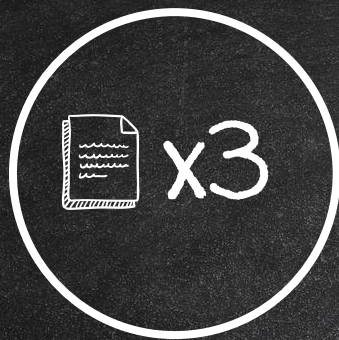
Michener 2015



Create 3 copies
(1 primary, 2 back ups)

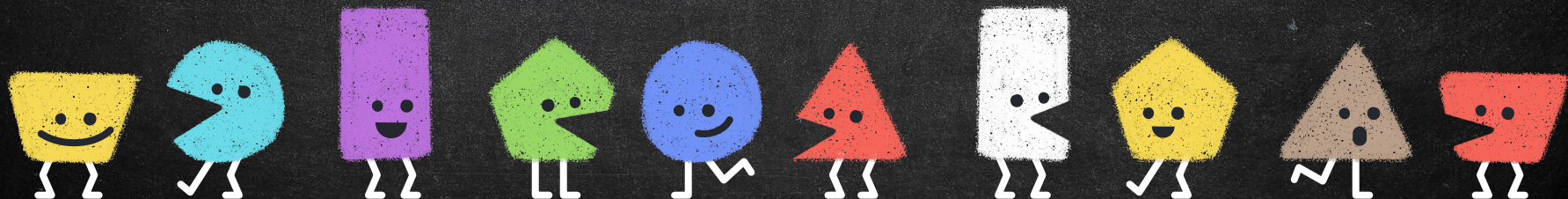
Use 2 storage media
(local or network
drive/NAS/Cloud)

Store 1 copy offsite
(fires, natural disasters)



Hart et al. 2016

BACK UP



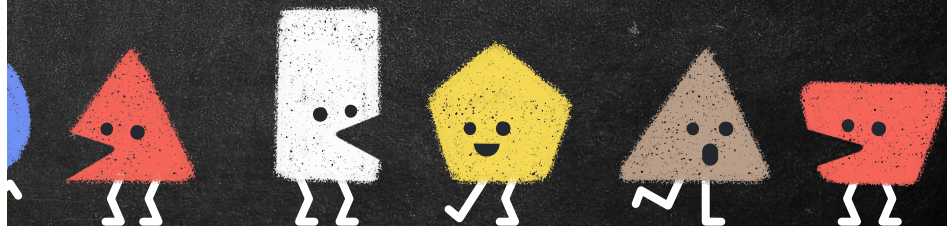
New addition to the 3-2-1 rule of backups:

Store your data on at least two distinct horizontal surfaces.

My adjustable height desk crashed this morning. My iMac faceplanted right onto the backup drive sitting on the same desk, and now both are toast. 🤖🤖🤖



BACK UP



VOCABULARY: HOW TO SAY FEMALE



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

METADATA

DATA ABOUT DATA

title, date, creator(s), keywords
describing the data

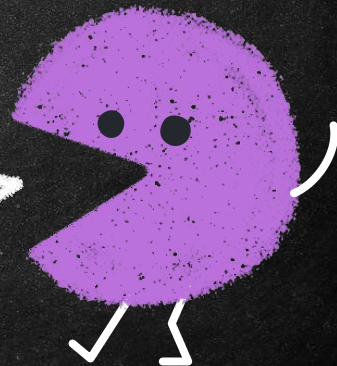
Metadata standards

- ensures interoperability and machine readability
- Use these resources to look for standards:

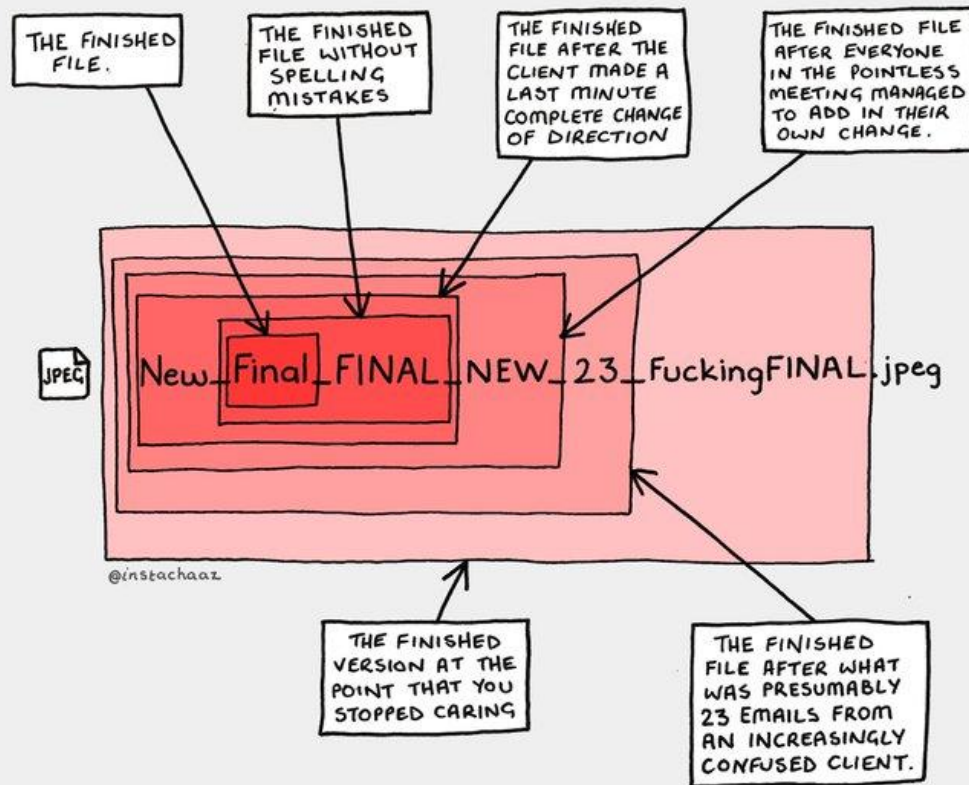
[FAIRsharing.org](https://fairsharing.org)

[Research Data Alliance metadata directory](#)

[Digital Curation Center](#)



ANATOMY OF A FILE NAME



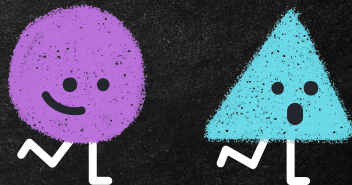
<https://twitter.com/chazhutton/status/1285955514241875968>

FILE NAMING

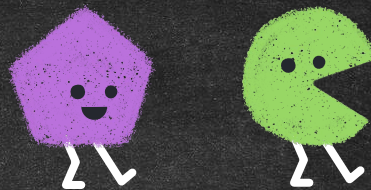
20210326-SeaChanges-RDM-V001

- Date or date range of experiment: YYYYMMDD (ISO 8601!)
- File type
- Researcher name/initials
- Version number of file (v001, v002)
- Don't make file names too long (30-70 characters should do the trick)
- Avoid special characters (? \ ! @ * % { [< >) and spaces

Check this [8 step guide](#) on how to set up your file naming convention or [this presentation on file naming](#), or [Stanford's best practices](#)



ORGANISATION



Create a consistent folder structure

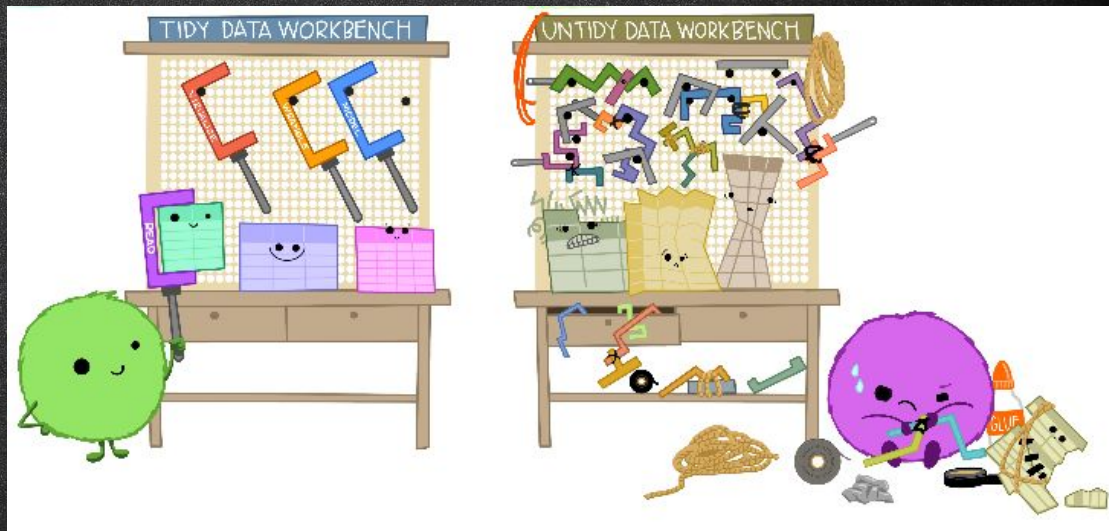
Folder structure templates

- [Colomb et al.](#)
- [Nikola](#)

[Find Files Faster: How to Organize Files and Folders](#)

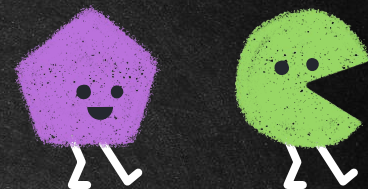
[Data Management: File organisation by Christine Malinowski](#)

[Videos on project structure](#) by Danielle Navarro



ORGANISATION

Create a consistent folder structure



Use the [cookiecutter](#) templates to download folder structures through **GitHub**. The template by Barbara Vreede follows the template proposed by [Wilson et al. \(2017\)](#): <https://github.com/bvreede/good-enough-project>

R: after installing cookie cutter,
“cookiecutter
gh:bvreede/good-enough-project”

```
.
├── .gitignore
├── CITATION.md
├── LICENSE.md
├── README.md
├── requirements.txt
├── bin
│   └── external
├── config
├── data
│   ├── processed
│   ├── raw
│   └── temp
├── docs
│   ├── manuscript
│   └── reports
├── results
│   ├── figures
│   └── output
└── src
```

- <- Compiled and external code, ignored by git (PG)
- <- Any external source code, ignored by git (RO)
- <- Configuration files (HW)
- <- All project data, ignored by git
- <- The final, canonical data sets for modeling. (PG)
- <- The original, immutable data dump. (RO)
- <- Intermediate data that has been transformed. (PG)
- <- Documentation notebook for users (HW)
- <- Manuscript source, e.g., LaTeX, Markdown, etc. (HW)
- <- Other project reports and notebooks (e.g. Jupyter, .Rmd) (HW)
- <- Figures for the manuscript or reports (PG)
- <- Other output for the manuscript or reports (PG)
- <- Source code for this project (HW)

SPREADSHEETS



Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Sometimes it's easier to rewrite genetics than update Excel

By **James Vincent** | Aug 6, 2020, 8:44am EDT

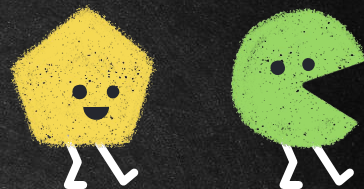
<https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates>



SPREADSHEETS

Create spreadsheets that are less error-prone, easier for computers to process, and easier to share

- Be consistent
- Write dates like YYYYMMDD
- Do not leave empty values (use NA)
- Put as few information possible in a single cell
- Create a data dictionary that describes the spreadsheet
- Leave the raw data alone
- Do not use colours
- Use data validation to avoid errors ([OpenRefine](#))



[Extended version](#)

[Broman and Woo](#)
[2018](#)

[Carpentries](#)
[Spreadsheets for](#)
[Ecologists](#) or [Social](#)
[Scientists](#)

TIDY DATA (R)

Tidy Data Paper by Wickham (see here for the CRAN code heavy version)

Welcome to the Tidyverse

The tidyverse style guide



“**TIDY DATA** is a standard way of mapping the meaning of a dataset to its structure.”

—HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

each column a variable

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

each row an observation

Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

<https://www.openscapes.org/blog/2020/10/12/tidy-data/> Illustration by Allison Horst (CC-BY)

VERSIONING



<https://twitter.com/electricarchaeo/status/1264362337559023616>

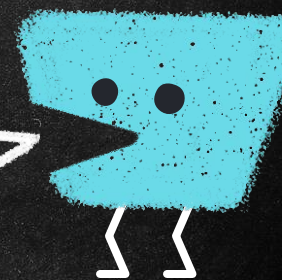
Versioning

who did what and
when

in collaborative work it
is useful to keep track of
different versions

Examples

- Google Docs/Drive
- Sharepoint
- Git (Hub)
- Version number of file (v001, v002)





Justin Stewart

@thecrobe

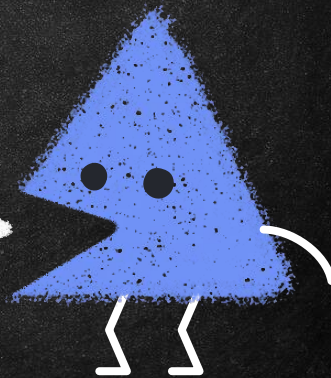
skimmed the protocol



<https://twitter.com/thecrobe/status/1373590641012322306>

DOCUMENTATION IS KEY!

- README files
- (Electronic) Labbooks
- Guide for data documentation
- Data Dictionary
- Code Book



GOOGLEDOC

Collaborative notes
that we're using for
questions/feedback
and the breakout
rooms!

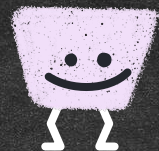


Data Management Workshop

26 March 2021

Agenda (in CET):

14:00	Intro Data Management
14:15	Icebreaker / Reflections assignments
14:35	🗨 Break out room 1
14:50	Reflection break out room 1
15:00	Break



BREAKOUT 1

Discuss your README template with someone from your discipline

Breakout room 1

1. Say hi and discuss one thing from this week that you learned or enjoyed
2. Share your README templates with each other through screen share or by reading part of the text to the person(s) you're grouped with:
 - Are you able to understand everything in the README file?
 - Is all of the information helpful?
 - Would you be able to (re)use the dataset without communicating with the creator of the data/README file: if not, what information are you missing?
 - What is the most important information about a dataset you would need in order to work with it?

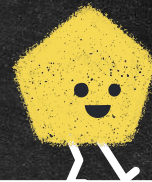
Optional notes:

Room 1:

-
-
-

BREAKOUT 2

Discuss your README template with someone from outside of your discipline



Breakout room 2

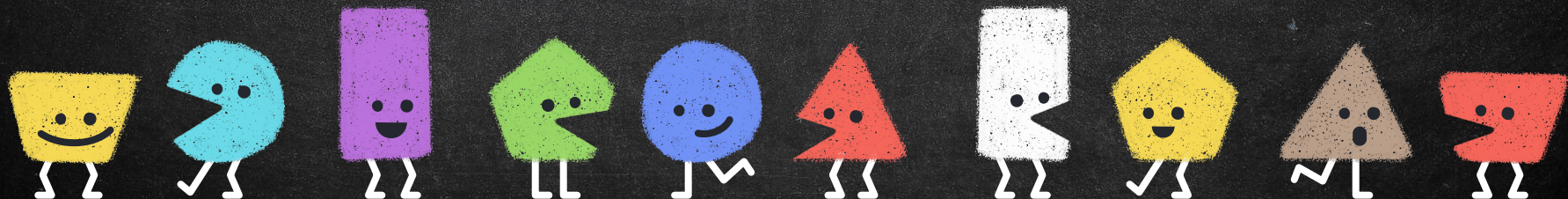
1. Say hi and discuss one thing from this week that you learned or enjoyed
2. Share your README templates with each other through screen share or by reading part of the text to the person(s) you're grouped with:
 - Are you able to understand everything in the README file?
 - Is all of the information helpful?
 - Would you be able to (re)use the dataset without communicating with the creator of the data/README file; if not, what information are you missing?
 - What is the most important information about a dataset you would need in order to work with it?

Optional notes:

Room 1:

-
-
-

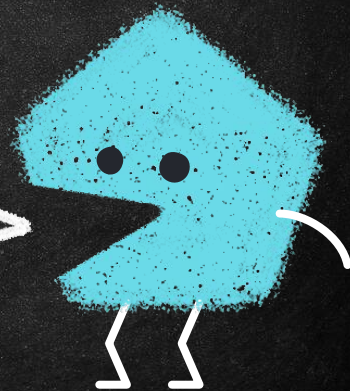
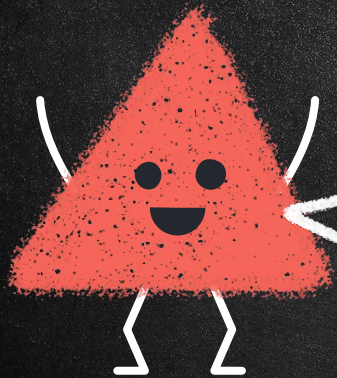
INTRO OPEN DATA



DATA OWNERSHIP

Who has the **rights** to your
project's data?

Are you allowed to share it?



OPEN DATA: MADE FREELY AVAILABLE FOR USE AND RE-USE BY ANYONE AND EVERYONE

Access

Available (on the internet)
to all on demand

Reuse/distribution

Provided under terms that
permit reuse and
redistribution

Transparency

Providing information about
data generation/collection

Interoperability

Interoperability with other
data, machine readable
formats

Participation

Everyone must be able to
use, reuse and redistribute

Equity

Data is not truly open if the
research process is not
open to all.

#bropenscience is broken
science by Kirstie
Whitaker and Olivia Guest

Open Science Beyond
Open Access: For and
with communities
A step towards the
decolonization of
knowledge



NOT
OPEN
DATA



MadScientist
@MadS100tist

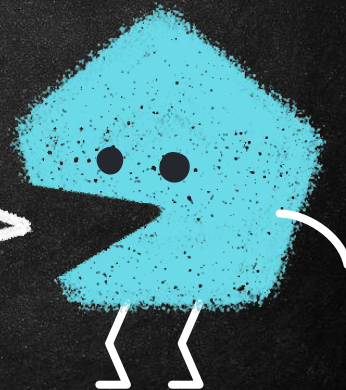
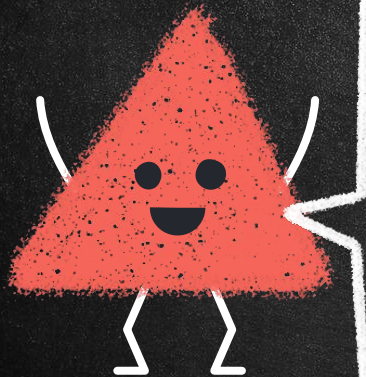
"Data will be available upon request"

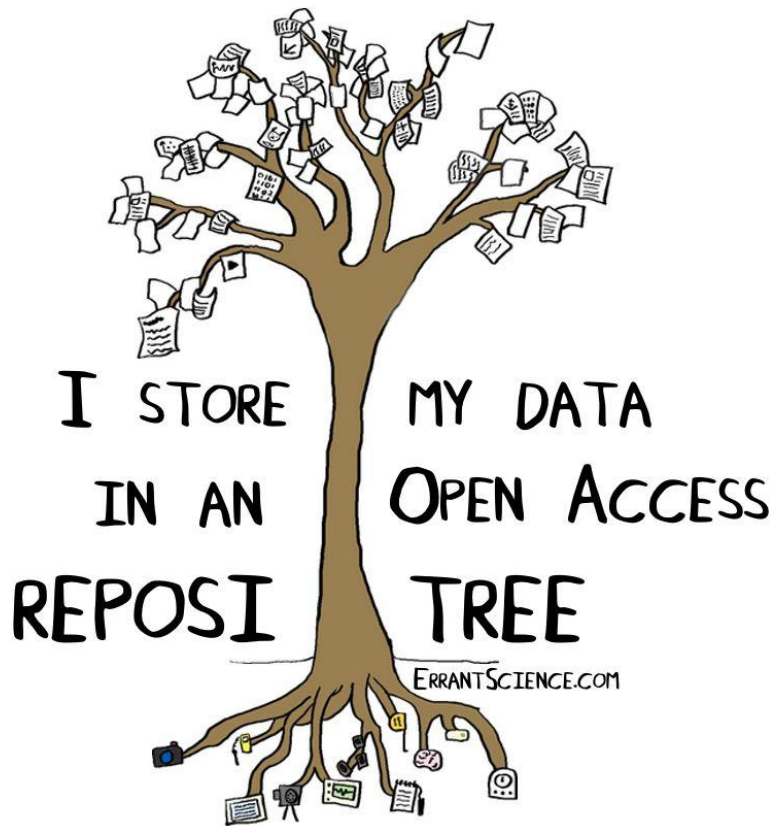


<https://twitter.com/MadS100tist/status/1366103674989277185>
<https://knowyourmeme.com/memes/agnes-harkness-winking>

*'research data
cannot be
reliably
preserved by
individual
researchers'*

Vines et al. 2014





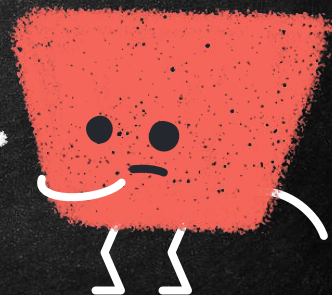
<https://twitter.com/ErrantScience/status/1251118457279647746/photo/1>

DATA REPOSITORY

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

- Finalised datasets
- ~10-15 years

How can you make research
data accessible? by Esther
Plomp



FAIR

[FAIR?](#) (Presentation slides)

29

[GO FAIR](#) overview

[Wilkinson et al. 2016](#) (Paper on FAIR)

[The Turing Way: FAIR](#) (general overview)

Findable

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

F

Accessible

- Determine what should be shared
- Ensure there is an access procedure in place
 - ≠ open!

A

- Use open/common formats and languages
- Consistent vocabulary
- Metadata standards

I

Interoperable

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

R

Reusable

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



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

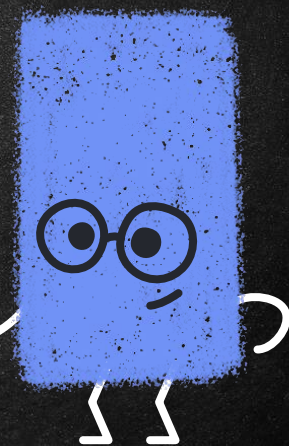


You can choose license
type for your adaptations
of the work.

CC-BY-SA <https://foter.com/blog/how-to-attribute-creative-commons-photos/>

Data:
Creative Commons
License Chooser

Software:
Choose an open
source license



PERSISTENT IDENTIFIERS

long-lasting reference
to a file, web page, or
other object

Preprints, articles, data,
software, persons,
samples

Citation!



ORCID
<https://orcid.org/>

iD

ORCID
Connecting Research and Researchers

ABOUT FOR RESEARCHERS MEMBERSHIP DOCUMENTATION RESOURCES NEWS & EVENTS

Esther Plomp
ORCID iD
<https://orcid.org/0000-0003-3629-1387>

First view

Websites & Social Links
Twitter
LinkedIn
Delft University of Technology profile page

Country
Netherlands

Keywords
Mass Spectrometry, Human Provenancing, Human Osteology, Biological Anthropology, Archaeology, Forensic, Neodymium, Strontium, Isotope, Biogeochemistry, Open Science, Research Data Management, RDM, Data Management Plan, DAM

Employment (2)
Data Science Journal
2020-12-04 | journal article
DOI: 10.3384/doi.2020-0458
Part of ISSN: 1683-1470

Education and qualifications (1)
Delft University of Technology
2019-08-25 | journal article
DOI: 10.3384/doi.2019-0458
Part of ISSN: 1683-1470

Membership and service (1)
Delft University of Technology
2019-08-25 | journal article
DOI: 10.3384/doi.2019-0458
Part of ISSN: 1683-1470

Funding (1)
Delft University of Technology
2019-08-25 | journal article
DOI: 10.3384/doi.2019-0458
Part of ISSN: 1683-1470

Works (58 of 58)

Items per page: 50 1 - 58 of 58

Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments
Data Science Journal
2020-12-04 | journal article
DOI: 10.3384/doi.2020-0458
Part of ISSN: 1683-1470
Source: Esther Plomp

Research Data Management
2020-09-24 | lecture speech
DOI: 10.3384/doi.2020-0458
Source: Esther Plomp

Strontium, oxygen, and carbon isotope variation in modern human dental enamel
American Journal of Physical Anthropology
2019-08-25 | journal article
DOI: 10.1002/ajpa.24050
Source: Crossref

FSC12020 Lightning Talk: The Turing Way
Zenodo
2019-08-11 | lecture speech
DOI: 10.3384/doi.2019-0458
Source: Esther Plomp

Citation: Briney KA, Coates H, Goben A (2020) Foundational Practices of Research Data Management. Research Ideas and Outcomes 6: e56508. <https://doi.org/10.3897/rio.6.e56508>

ADDITIONAL RESOURCES

Turing Way

The Turing Way is an open source community-driven guide to reproducible, ethical, inclusive and collaborative data science.

Noble 2009

A Quick Guide to Organizing Computational Biology Projects

Borer et al. 2009

Some Simple Guidelines for Effective Data Management

Briney 2015

Data Management for Researchers

Kramer & Bosman

Rainbow of open science practices

Ainsworth 2019

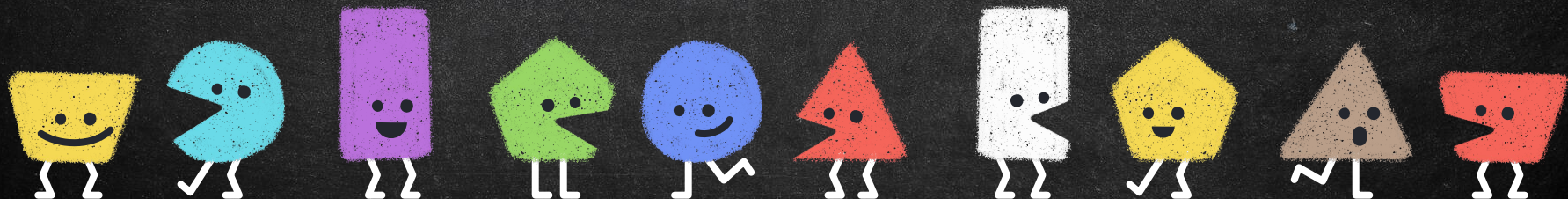
Research Culture is Broken: Open Science can Fix It (video)



THANKS!

Any questions? -> GoogleDoc
Reflection/Feedback

You can find me at @PhDToothFAIRy &
e.plomp@tudelft.nl



CREDITS

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by [SlidesCarnival](#)
- Photographs by [Unsplash](#)

