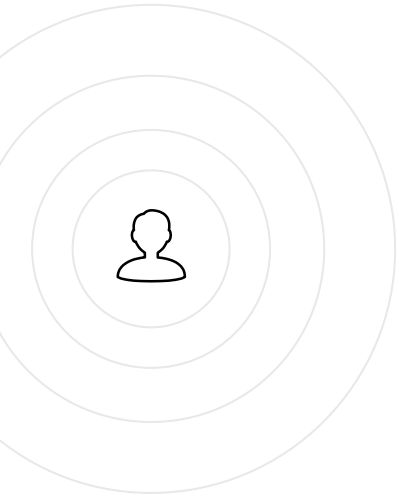


# Data Management Plans, why and how?

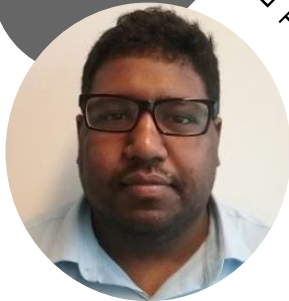
Santosh Ilamparuthi & Esther Plomp  
Helis Academy 27th of May

[https://www.doi.org/  
10.5281/zenodo.3167795](https://www.doi.org/10.5281/zenodo.3167795)



# Santosh Ilamparuthi

Data Steward  
TU Delft, Faculty of Electrical Engineering  
Mathematics and Computer Science  
s.ilamparuthi@tudelft.nl



# Esther Plomp

Data Steward  
TU Delft, Faculty of Applied Sciences  
e.plomp@tudelft.nl



# Research Data Management

- The **organisation of data** throughout the research project
- **everyday management** of research data (storage, file naming, preservation)
  - how will data be **preserved and shared** after the project is completed?

# Research Data Management

Prevent data loss

Unfindable files

Can your data be reused by others when you leave?

Recognition for all research outputs

Collaboration

Increases quality of scientific practice

Cost/time efficient

# Data Loss - What if?

## THE FOUR STAGES OF DATA LOSS DEALING WITH ACCIDENTAL DELETION OF MONTHS OF HARD-EARNED DATA



# Data Loss – What if?



Iestyn Shapey  
@iestyn\_shapey

Volgen

Could the individual who has just stolen my laptop (which contains the PhD I was due to submit in 2-3 weeks time) from a secure office @MFTnhs please return it immediately! I don't care about the computer, but my work is irreplaceable & has the potential to transform many lives.

Tweet vertalen

16:53 - 25 mei 2019

2.070 retweets 888 vind-ik-leuks



Sam Giles  
@GilesPalaeoLab

Volgen

Hey #AcademicTwitter, how often do you remember to back up your laptop?

Tweet vertalen

19% Every day

18% Every week

32% Errr every few months?

31% I bought a drive once...

189 stemmen • 22 uur resterend

17:39 - 26 mei 2019

16 2 3

# Data Loss – What if?

- Your laptop/notebooks got stolen/lost?
- Your workplace/lab burnt down?
- You lost your USB stick?
- Your external drives are damaged?
- Your files on Dropbox/Google drive disappeared?





# Read the small print!

## Google services Terms of Use:

When you upload, submit, store, send or receive content to or through our Services, you give Google (and those we work with) a worldwide license to use, host, store, reproduce, modify, create derivative works (such as those resulting from translations, adaptations or other changes we make so that your content works better with our Services), communicate, publish, publicly perform, publicly display and distribute such content. The rights you grant in this license are for the limited purpose of operating, promoting, and improving our Services, and to develop new ones. This license continues even if you stop using our Services (for example, for a business listing you have added to

<https://www.google.com/intl/en/policies/terms/>



**F**indable 

**A**ccessible 

**I**nteroperable 

**R**eusable 

# Findable

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



**F**

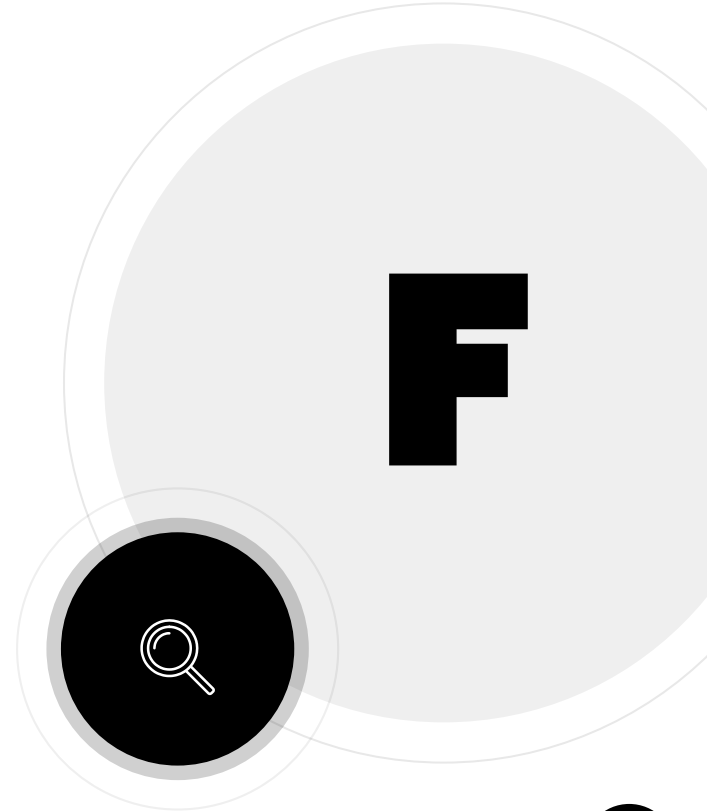


# Findable

- 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



# 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?

-> reproducible research

A large, bold, black letter 'F' is centered within a light gray circular area. This area is part of a larger graphic on the right side of the slide, which includes a magnifying glass icon in a smaller circle below it.

# create useful README files

```
Cornell AUTHOR_DATASET_ReadmeTemplate.txt

This DATSETNAMEreadme.txt file was generated on [YYYYMMDD] by [Name]

-----
GENERAL INFORMATION
-----

1. Title of Dataset

2. Author Information

Principal Investigator Contact Information
  Name:
  Institution:
  Address:
  Email:
```

<https://data.research.cornell.edu/content/readme>  
README files template:  
<https://cornell.app.box.com/v/ReadmeTemplate>

Original slide by Marta Teperek



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

E. Plomp <sup>a</sup>, I.C.C. von Holstein <sup>a</sup>, J.M. Koornneef <sup>a</sup>, R.J. Smeets <sup>a</sup>, J.A. Baart <sup>b, c, 1</sup>, T. Forouzanfar <sup>b, c</sup>, G.R. Davies <sup>a</sup>

Show more

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

Under a Creative Commons license

Get rights and content

open access

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

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

ORCID



# Accessible

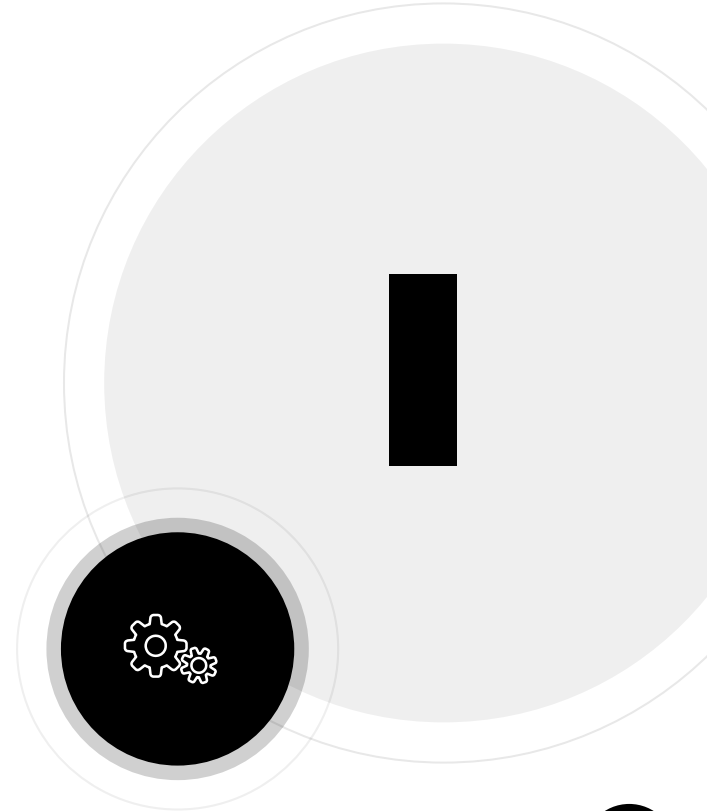
- Consider what will be shared
- Obtain participant consent and perform risk management
- Determine access control

**A**



# Interoperable

- Use open/common format/languages
- Consistent vocabulary
- Discipline common metadata standards
  - [FAIRsharing.org](https://www.fairsharing.org)
  - [Research Data Alliance metadata directory](#)
  - [Digital Curation Center](#)





# Reusable

- Apply a licence
- Documentation

**R**



# Reusable

- Apply a **licence**
- Documentation

<https://creativecommons.org/licenses/>

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

## 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 Licence

Apache Licence 2

GNU General Public Licence 3 (GNU GPLv3)

A large, bold, black letter 'R' is positioned on the right side of the slide, partially overlapping a light gray circular graphic element.

**F**indable   
**A**ccessible   
**I**nteroperable   
**R**eusable 

<https://www.go-fair.org/fair-principles/>









**From FAIRlyle to reality:  
Research Data Management put into practice**

**F**indable

To identify your data to a data repository (an online archive that stores research datasets and provides long-term access to them), you ensure that your datasets are given a **persistent identifier** (a long-lasting reference). This persistent identifier can be a Digital Object Identifier (DOI), that allows others to find the location of your datasets and provides the ability to request when they cite your dataset in their publication. Data repositories use persistent identifiers to date using common identifiers, allowing search engines to find your datasets which increases the **visibility** and **impact** of your research outputs.

By connecting **metadata** (information about your dataset) to your identifiers, others are able to find it through keywords or author information.

Examples of useful data repositories are:
 

-  Zenodo
-  Open Access
-  Data Commons
-  Dataverse
-  Pangea
-  ORCID

By making use of **GitHub** and **Research** you can get a persistent identifier for your code.

You can receive a DOI for your research projects or products.

You can also create a persistent identifier for yourself by creating an **ORCID**.

**A**ccessible

Before uploading your data, protocols and software to a repository, you must decide which research outputs you will make available to others. Select the relevant data and code and remove unnecessary data before uploading to clear the selection you will have to determine how others will be able to access your research outputs: can anyone access them (**Open Access**) or should they request access (**restricted access**)? By uploading your data, protocols and software to a repository you can define who has access there.

Accessibility does not equal **Open**! Instead the conditions under which the research outputs are accessible should be specified. If you cannot make your data publicly available (**personal/confidential data**), it is possible to archive your data under **restricted access** to a data repository, such as **DAZ**.

Even when your data is not openly accessible, or no longer available, the metadata should be persistently available to indicate that the datasets exist.

**I**nteroperable

To make your research output **human and machine readable**, you should make use of **standardized data standards** or **terminologies**, controlled vocabularies and ontologies or **computational languages** frequently used within your research community.

Link your data, protocols and software to related research outputs by **ORCID**, these outputs or sharing from your research outputs are derived from them. This also provides others with credit for their work.

Document the **provenance** of your code or scripts such that other researchers can execute them easily in other environments.











Prefer **open data formats** (.xml, .csv), over proprietary formats (.xls).










**R**eusable

Document your data and software in a way that humans and machines can understand it. This includes all the information that is required to evaluate your research outputs that was the context in which it was generated. It refers to the data generated, code, and software/protocols/ software used. To allow the reuse of **DATA** flow use the data generated. Use a standardized way of providing the information and organizing your research outputs, for example, writing a **README** text file that accompanies your datasets and code. By setting up a **Data Management Plan (DMP)** you can plan how to effectively manage and archive your research outputs.

Repositories will allow you to select a license for your data and software. A license is a standardized machine-readable statement that tells the user under which conditions they can use the data. Some examples of licenses are: a **CC BY** license you will allow others to distribute and build upon your work, as long as they credit you for the original creation. Using the **CC BY** license for software allows others to reuse your code with the restriction, when you make data and software, always identify the license or conditions that inform the way you did use it.

The **FAIR** principles allow for the broadest possible use of your research outputs, by ensuring consistent collaborations, machine and the general public.

TU Delft         

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

**F**indable 

**A**ccessible 

**I**nteroperable 

**R**eusable 

≠

**Standard**

**Open**

**Intrinsic  
quality**



Current Issue



[JCB Home](#) > [2013 Archive](#) > [28 October](#) > [Mirouse et al. 203 \(2\): 373](#)

Published October 28, 2013 // JCB vol. 203 no. 2 373

[The Rockefeller University Press](#), doi: 10.1083/jcb.20070205310112013r

© 2013 Mirouse et al.



## Retraction

### LKB1 and AMPK maintain epithelial cell polarity under energetic stress

Vincent Mirouse, Lance L. Swick, Nevzat Kazgan, Daniel St Johnston, and Jay E. Brenman

Vol. 177 No. 3, April 30, 2007. Pages [387–392](#).

The editors of *The Journal of Cell Biology* have been notified by Dr. Daniel St Johnston and Dr. Jay E. Brenman that they and the other authors of the paper referenced above retract the paper. As a result of this retraction, no data in this paper should be cited in the scientific literature.

#### Views

[Retraction to Mirouse et al. 177 \(3\): 387](#)

[» Full Text \(HTML\)](#)

[▼ Top](#)

[▼ References](#)

[Full Text \(PDF\)](#)

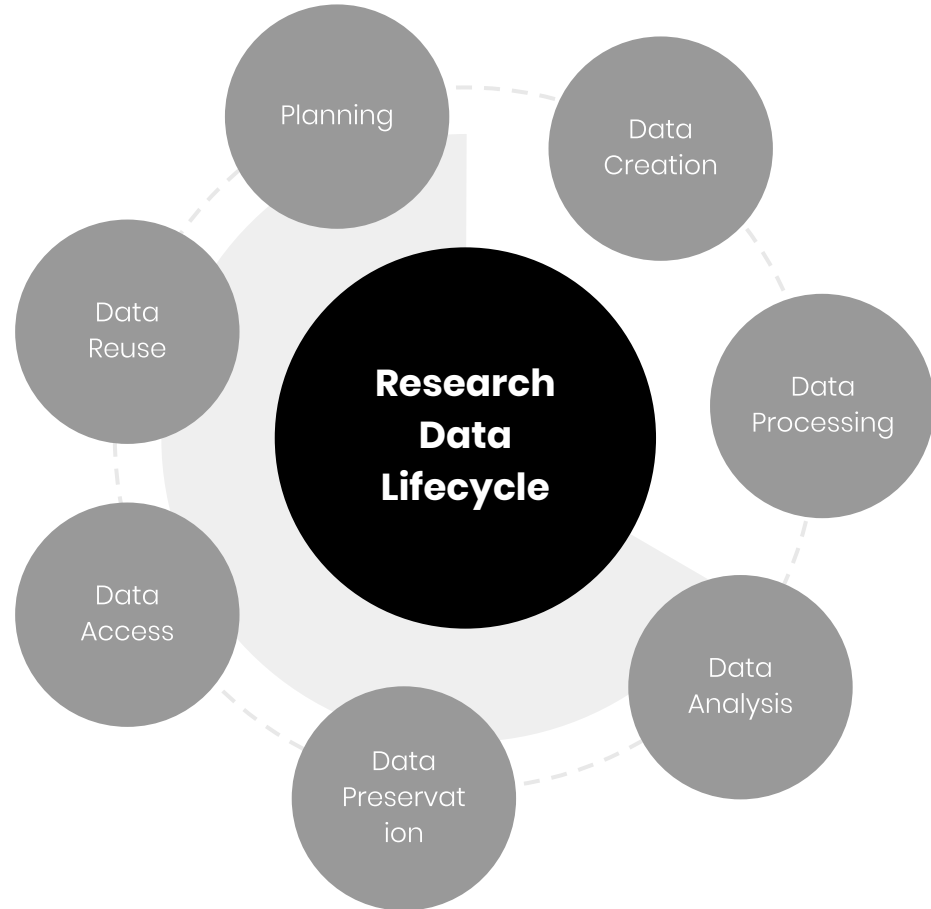
[Article Usage Statistics](#)



# DMP

- **Planning how to manage your data during the research data life cycle**

# Research Data Life Cycle





# Data Management Plan

## Data

Format

Raw

Processed

Final

Organisation  
Strategy

Standards

Personal Data?

## Storage

Size

Back up (two  
locations)

Costs (during and  
after project)

Access

## Preservation

Sharing?

Available

Persistent Identifier

## Reuse

Licence

Patents

Personal data



# Funder requirements

## NWO

Data Paragraphs

Data Management Plans

[More info](#)

## ZonMw

Data Management Plans

Tools

[More info](#)

## European Commission

H2020, Horizon Europe

Data Management Plan

Policy

[More info](#)

## Hartstichting

Data Management Plan

[More info](#)



# Data Organisation

- Consistent
- Meaningful to you and your colleagues
- Allow you to find files easily

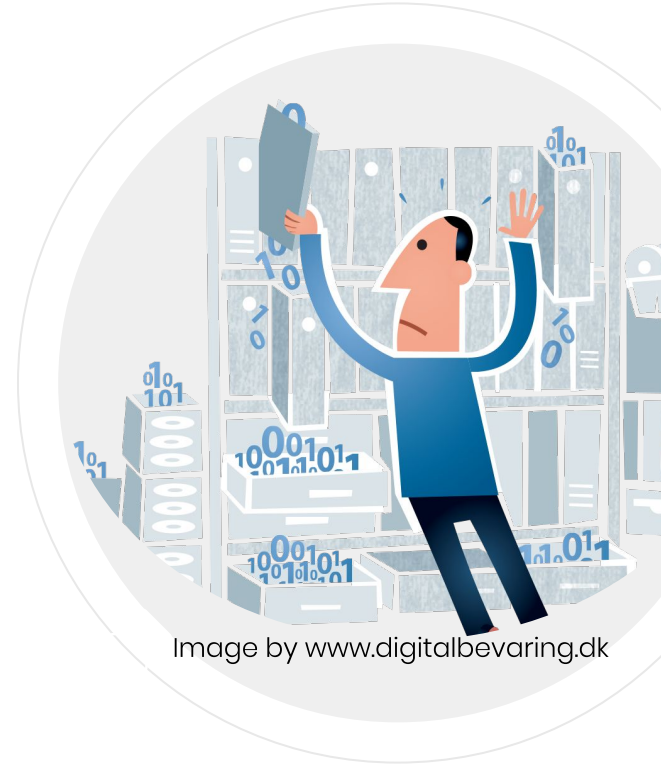
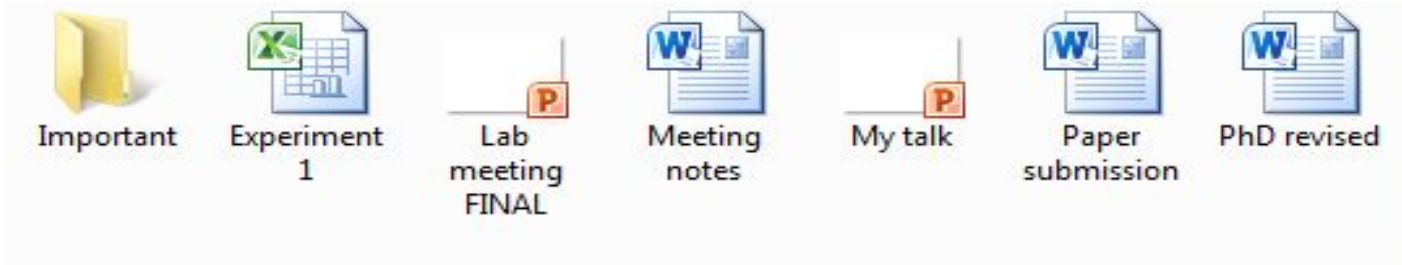


Image by [www.digitalbevaring.dk](http://www.digitalbevaring.dk)

# Data Organisation

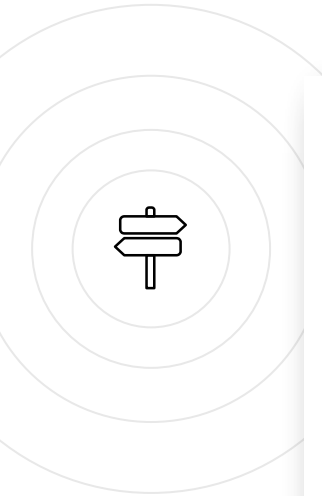


In 3 years time would you know what these are?

# Data Organisation



Copyright: <http://10pm.com/>



## Example A

### Documents library

PhD data

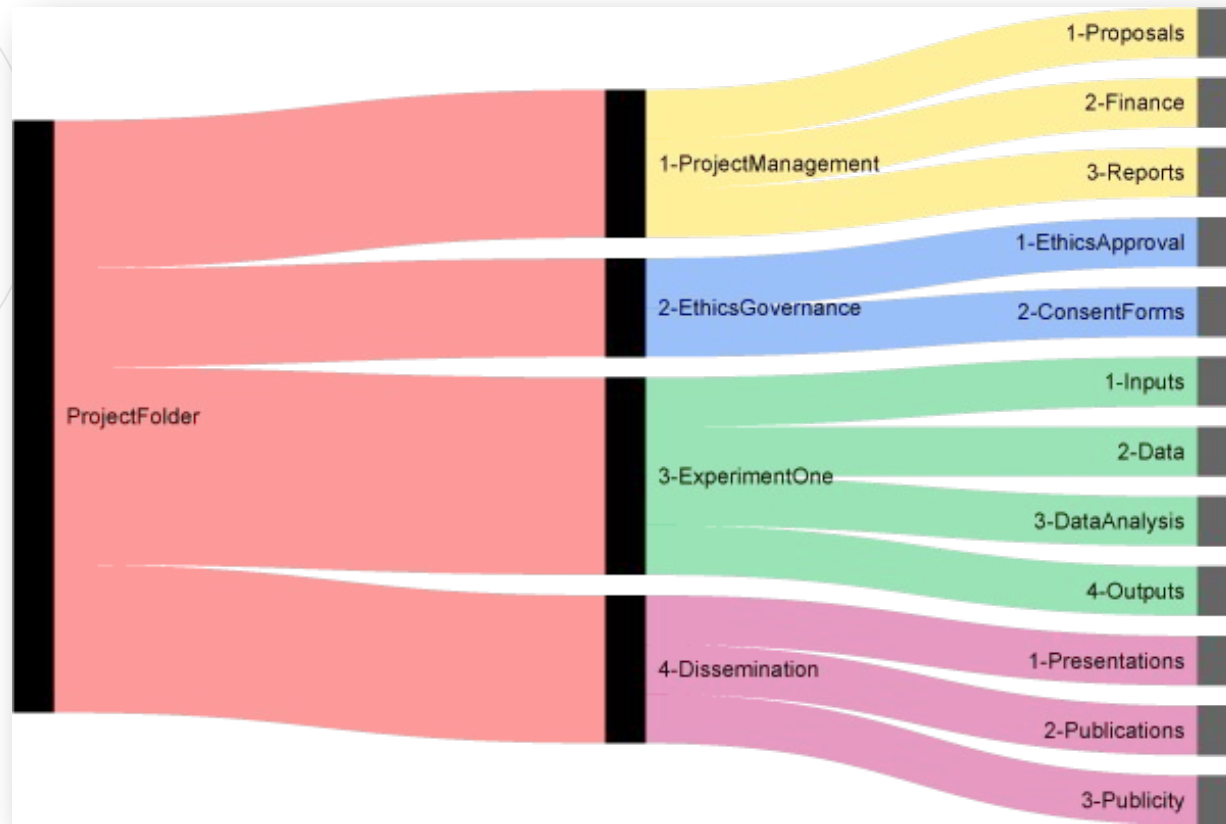
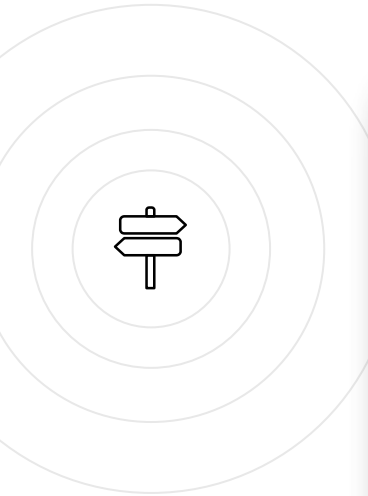
- 25July
- Documents
- Experiments
- Experiments2
- From desktop
- Important
- Other
- PhD
- Talks
- Experiment 1
- Lab meeting FINAL
- Meeting notes
- My talk
- Paper submission
- PhD revised

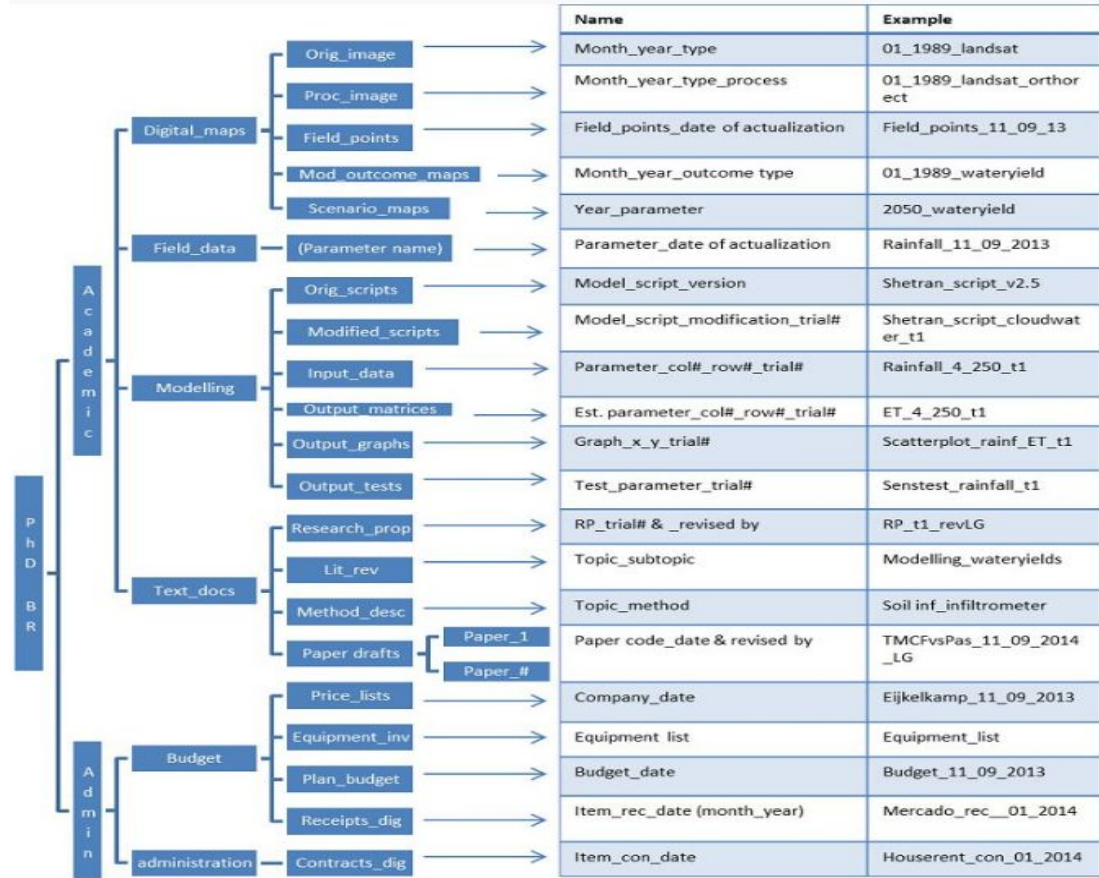
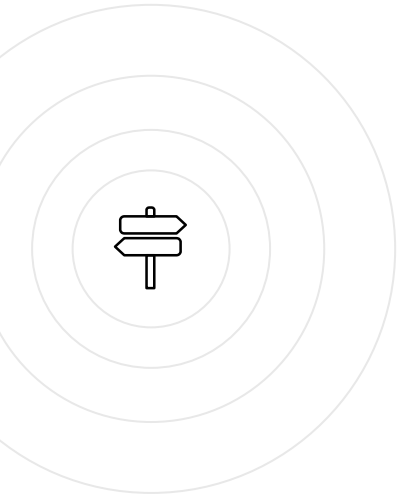
## Example B

### Documents library

PhD data2

- Conferences
- Downloaded publications
- Experimental data
- Financial documents
- PhD thesis
- Presentations
- Protocols
- Reagents
- Reports
- Training





# File Naming Conventions

## **20190527\_HelisAcademy**

- Date or date range of experiment: YYYYMMDD
- File type
- Researcher name/initials
- Version number of file
- Don't make file names too long
- Avoid special characters and spaces
- Include a README.txt file to explain the naming convention



# Data Organisation

Reference your samples:

- dates in notebooks + supplier's name/code

Add any relevant notes

	A	B	C	D	E
1	<b>Box #1</b>				
2	<b>Name of the sample</b>	<b>date in lab book</b>	<b>Samples left (if multiple samples)</b>	<b>Sample size</b>	<b>Notes</b>
3					
4	Fibroblasts protein extract	23/03/2014	1	~40ul	Concentration: 0.6mg/ml
5	Fibroblasts protein extract	05/04/2014	7	~40ul	Concentration: 0.65mg/ml Difficulties measuring concentration
6	Fibroblast protein extract for mass spec	19/08/2014	4	~50ul	Concentration: 80ng/ul
7	Fibroblast RNA extract	20/09/2014	12	~10ul	Concentration: 100ng/ul; excellent prep
8	Fibroblast RNA extract	23/09/2014	2	~10 ul	Abcam: ab8227 LOT: GR47300
9	Anti-actin antibody	N/A	18	~5ul	Concentration: 0.6mg/ml Rabbit polyclonal IgG
10					
11					
12					



Vincent Gaggioli



# Version control

- Git
- Subversion
- Electronic Lab Notebooks



Commits on May 6, 2019

Update \_config.yml

 EstherPlomp committed 9 days ago ✓

Verified



5d7c79c

Update index.md

 EstherPlomp committed 9 days ago ✓

Verified



8d9f3c6



Commits on Apr 17, 2019

Update index.md

 EstherPlomp committed 28 days ago ✓

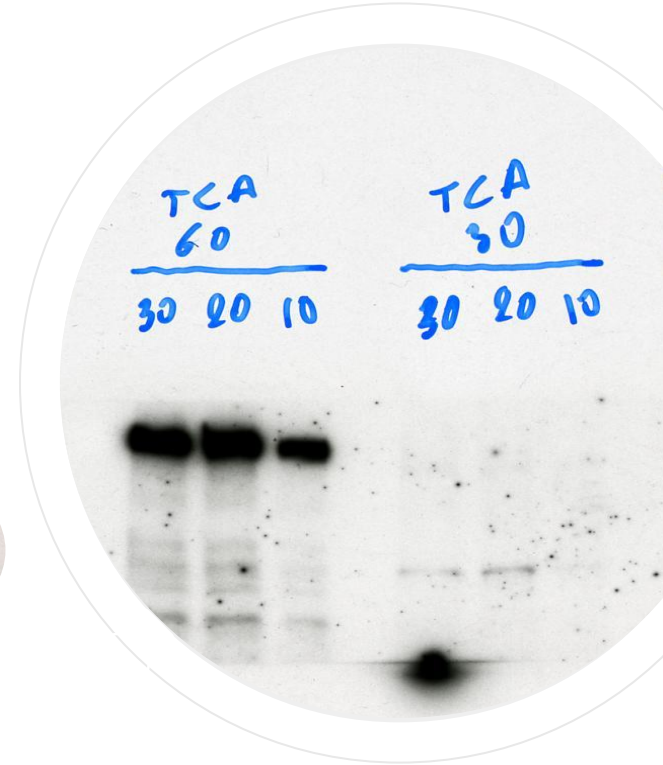
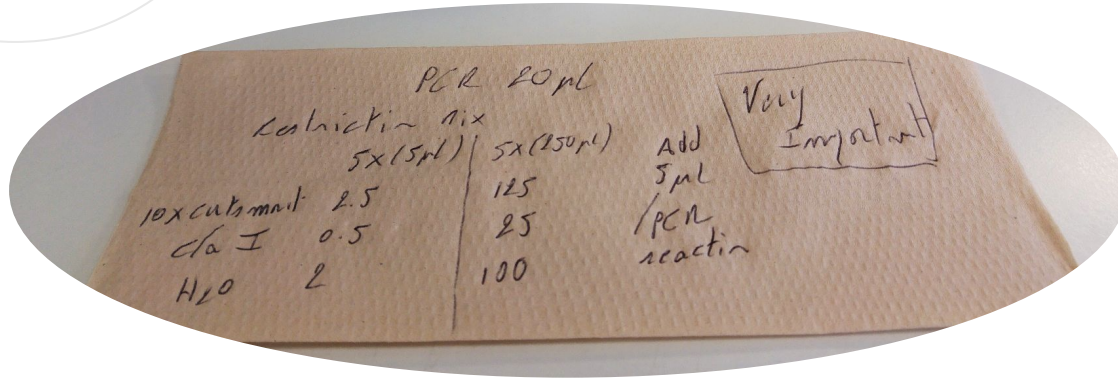
Verified



0b86713



# Data Documentation

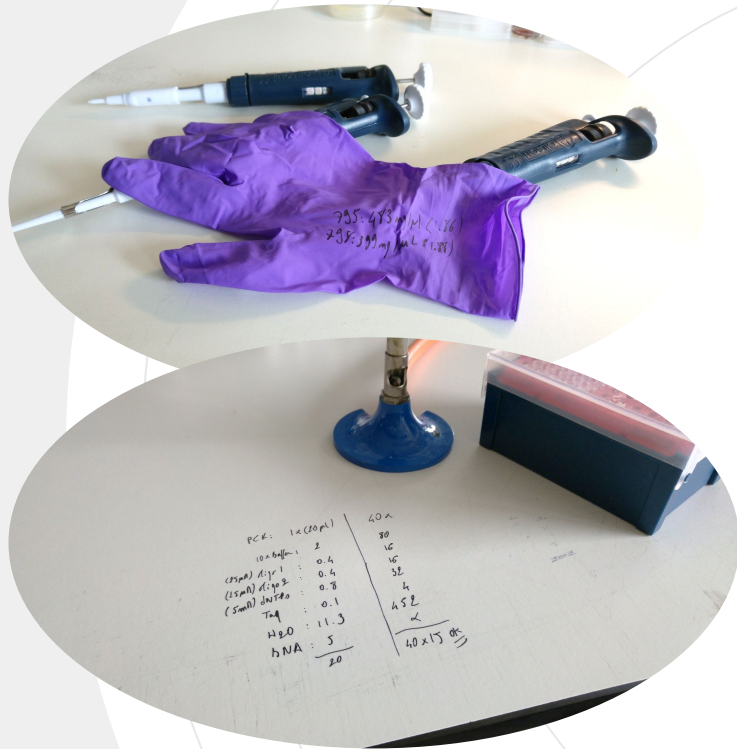


# Lab Notebooks

Documentation of experiments

- from hypothesis to results
- crucial for reproducibility and reuse of research

Discipline/Group/Individual specific



# Lab Notebooks



- Not searchable
- Handwritten
- Not reusable
- No direct link to (digital) data
- Difficult to back-up

**decrease research  
efficiency/  
reproducibility**

# Electronic Lab Notebooks

- Searchable
- Readable
- Reusable
- Direct link to (digital) data



More info on ELNs:  
<https://doi.org/10.5281/zenodo.2634449>



**increase research  
efficiency/  
reproducibility**



# Preregistration

- **Planning your data collection**
- **[Open Science Framework](#)**
- **[As Predicted](#)**
- **[Center for Open Science](#)**

# Data Repositories

Finalised Datasets  
Snapshot

Long-term  
preservation  
~10-15 years

## Findable

DOI  
Metadata

## Accessible

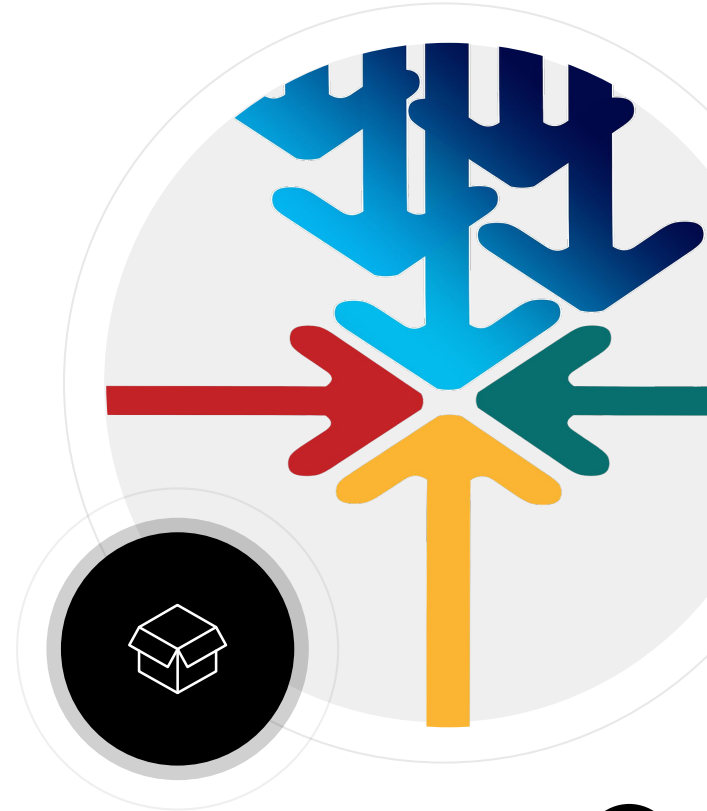
Control

## Interoperable

Metadata  
Vocabulary  
Open  
formats/standards

## Reusable

Licence





# Repositories



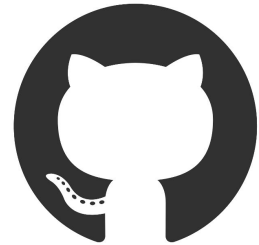
4TU.Centre for Research Data

European Genome-phenome Archive



[Recommended Repositories](#) (nature)  
[Recommended Repositories](#) (springer nature)  
[Registry of Research Data Repositories](#)

# Sharing Platforms



# General Data Protection Regulation (GDPR)

In effect since 25 May 2018

Applies to:

- **any EU researcher** who collects personal data about a citizen of any country
- Any non-EU researcher collecting **personal data on EU citizens**
- 'personal data' from **living persons**

Anonymised or de-identified data is NOT personal data

Original slide by Veerle Van den Eynden  
<https://zenodo.org/record/1408108#.XGUQkTBKjIU>



# General Data Protection Regulation (GDPR)

Original slide by Veerle Van den Eynden  
<https://zenodo.org/record/1408108#.XGUQkTBKjIU>

## Grounds for processing personal data

- **Consent** of the data subject
- Necessary for the performance of a **contract**
- **Legal obligation** placed upon controller
- Necessary to **protect vital interests** of the data subject
- Carried out in the **public interest** or is in the exercise of official authority
- **Legitimate interest** pursued by controller



# During projects

Collect and store **only what is necessary**

**Restricted access** + encryption

- files/folders: VeraCrypt
- disks: BitLocker (Windows) / FileVault (Mac)

**Informed consent**

- also personal data and needs to be securely stored



# After projects

Fully de-identified data can be published  
in an open access repository

Anonymisation vs. pseudonymisation

Open informed consent?

If not it can be published in a restricted  
access repository ([EASY](#))





# Thanks!

**Any questions?**

2

# DMPonline

<http://dmponline.dcc.ac.uk/>





# Features

Templates of most major funders available

Organization specific templates can be added

Can be shared with other researchers

Guidelines as to how to fill in the DMP is provided

Data Stewards can directly review the DMP (TU Delft)

DMPs are private by default but can be publicly shared

DMP can be downloaded

# Making a DMP

Create an account on with your institute credentials,  
<http://dmponline.dcc.ac.uk/>

Provide details of project and select appropriate template

Follow guidelines to fill in the DMP

Choose with whom you would like to share the DMP and  
request a review by the Data Steward (TU Delft)



# Credits

- Special thanks to Marta Teperek for re-using her slides
- Presentation template by [SlidesCarnival](#)
- Photographs by [Unsplash](#) and Vincent Gaggioli

