

# *NUTRIOME workshop*

## *Research Data Management (RDM)*

Mijke Jetten, FAIR data lead/Community Manager Data Stewardship

[mijke.jetten@health-ri.nl](mailto:mijke.jetten@health-ri.nl)

Maastricht, May 31, 2024

## Programme

09:30-10:15 Introduction to RDM

10:15-11:00 Hands on - *Ten practices: 'prepare well to prevent data disaster'*

11:00-11:15 Coffee/tea

11:15-12:00 Hands on - *RDMkit: best practices & guidelines to help you make your data FAIR*

**Objectives:** by the end of this workshop, you will be able to...

- recognise the basics of research data management
- apply research data management solutions to your research project
- evaluate what actions need to be taken to solve research data management issues in the NUTRIOME research project.

For the exercises we use shared notes.

**Raise your hand if you have any questions!**

We've heard rumours about ...

*'data management & stewardship not being the first most exciting subjects you think of when considering a workshop'*



As an icebreaker, to get to know each other a bit better, use the **shared notes** to share with us:

- how you rather spend today, honestly ... (we love a humorous reply)
- what data management things you (or your project colleagues) are already good at (we know you can think of something; as we all know, small steps move mountains)

# The future of science is Open

- The practice of science so that others can collaborate and contribute,  
  
*... where research data, lab notes and other research processes are freely available,  
... under terms that enable reuse, redistribution and reproduction of the research,  
... and its underlying data and methods*
- It includes open access publishing, sharing data, sharing code and sharing workflows
- Opening research supports validation, reproducibility and reduces cases of academic misconduct

*... it helps to maximise the impact of your research and provides the foundations for others to build upon  
... applying open science in your daily workflows is just part of good research practice*

Old school	Modern
Academic	Society
Individual	Team
Possessing	Sharing
Elite / Hobby	Necessity
Slow	Quick
Excellence	Impact
...	...





# FAIR data stewardship/management

## Data stewardship

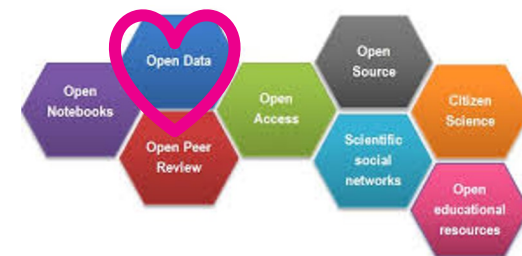
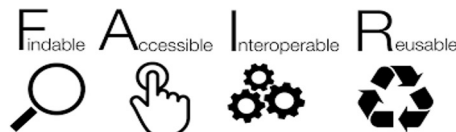
Responsible planning and executing of all actions on digital data before, during and after a research project, with the aim of optimizing the usability, reusability and reproducibility of the resulting data

## Long term view

Data stewardship focuses on the long term, stimulating researchers together with their data support staff (data stewards) to plan data in such a way that is maximises the reuse of data

## Data stewardship benefits the researcher

- Data is output! Data stewardship helps you to make conscious decisions about the data in your project
- Data stewardship prevents unauthorised access, avoids data loss, and facilitates the documentation and reuse of data
- Data stewardship stimulates efficiency and helps you comply to the conditions set to data by universities, funders and journals
- Data stewardship stimulates open science, FAIR data and helps meeting privacy- and security standards
- However, rule of thumb: *as open as possible, as closed as necessary*



## FAIR data principles

- **FAIR for human beings, for machines and taking into account legal requirements (such as privacy)**  
FAIR refers to data (or any digital object), metadata (information about that digital object), and infrastructure
- The first step in (re)using data is to making data **findable**. Metadata and data should be easy to find for both humans and computers. For that, rich and machine-readable metadata and persistent identifier are essential
- The data must be **accessible**. A description should exist of how the data can be obtained. And there should be guarantees that this will still work after years.
- The data should be **interoperable**. It should be structured in well described way, with standardised metadata and vocabularies. This way, it can be integrated with other data, and can operates with applications or workflows for analysis, storage, and processing.
- The ultimate goal of FAIR is to optimise the **reuse** of data. To achieve this, metadata and data should be well-described, including a clear license, so that they can be replicated and/or combined in different settings.

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



# Data life cycle



<https://www.jisc.ac.uk/guides/research-data-management-toolkit>

<https://ukdataservice.ac.uk/manage-data/lifecycle.aspx>

## Data life cycle: planning data



Design research; plan data management; plan consent for sharing; plan data collection, processing protocols and templates; explore existing data resources

### Topics

- Writing a data management plan
- Informed consent procedures

**You'll find a summary slide with basic information for each of these topics at the end of this presentation**



## Data life cycle: collecting data



Collect data; capture data with metadata; acquire existing third party data

### Topics

- Finding data
- Collecting data
- Preregistration
- Data security
- Personal data

**You'll find a summary slide with basic information for each of these topics at the end of this presentation**



## Data life cycle: processing and analysing data



Enter, digitize, transcribe and translate data; check, validate, clean, anonymize; derive data; describe; manage and store data; analyse and interpret data; produce research output; cite data sources

### Topics

- Data storage
- Data organisation
- Data versioning
- Data documentation

**You'll find a summary slide with basic information for each of these topics at the end of this presentation**



# Data life cycle: publishing, preserving & reusing data

PUBLISHING  
AND  
SHARING  
DATA

## Publishing and sharing data

Establish copyright; create user documentation; create discovery metadata; select appropriate access to data; publish/share data; promote data

PRESERVING  
DATA

## Preserving data

Migrate data to best format/media; store and backup data; create preservation documentation; preserve and curate data

RE-USING  
DATA

## Re-using data

Conduct secondary analyses; undertake follow-up research; conduct research reviews; scrutinize fundings; use data for teaching and learning

## Topics

- Data archiving
- FAIR software
- Data rights

**You'll find a summary slide with basic information for each of these topics at the end of this presentation**





# Ideal/theory vs. practice



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





# The data problem (reproducibility crisis)

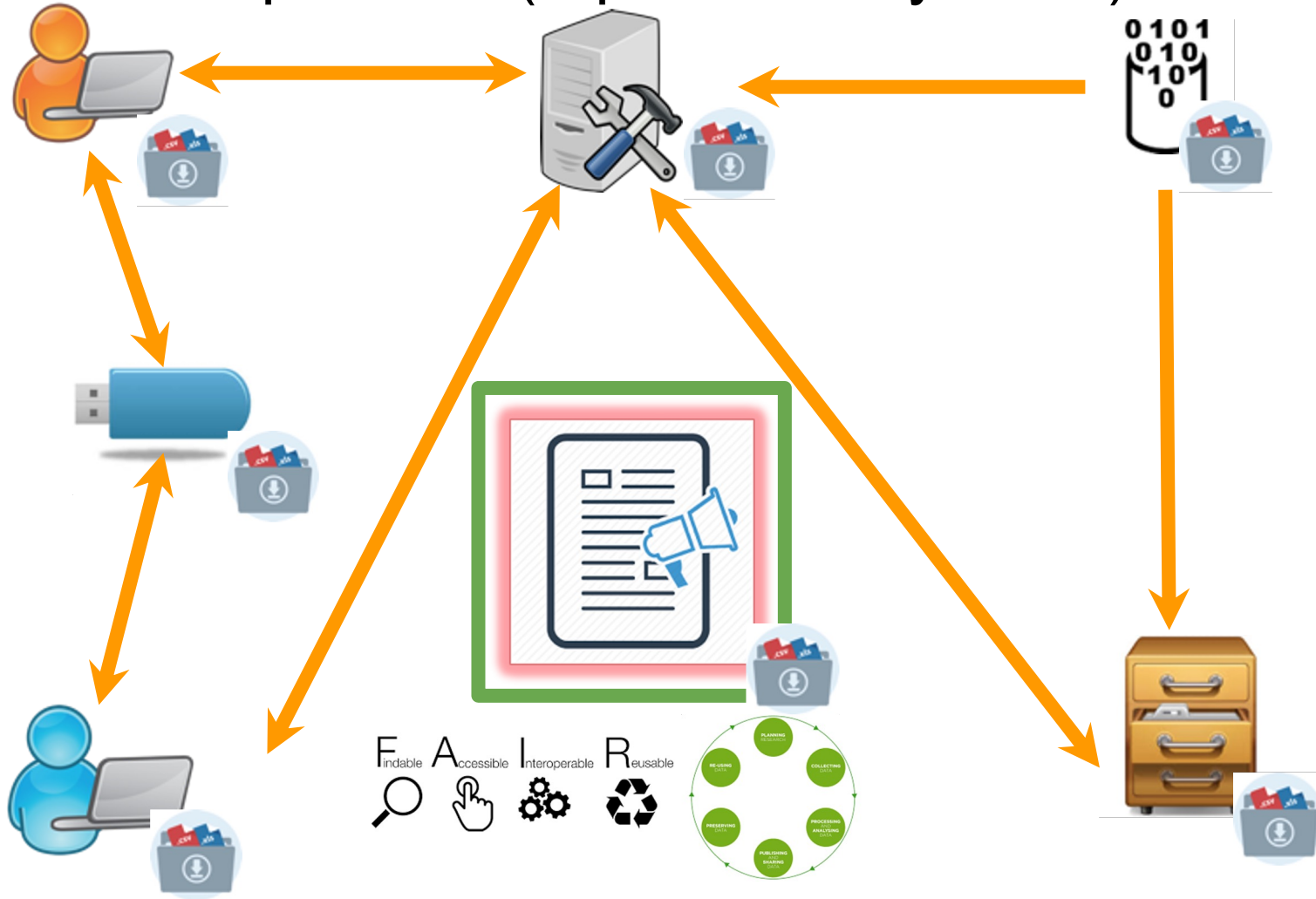


Image: Christine Staiger



# The data problem

<https://www.youtube.com/watch?v=N2zK3sAtr-4>



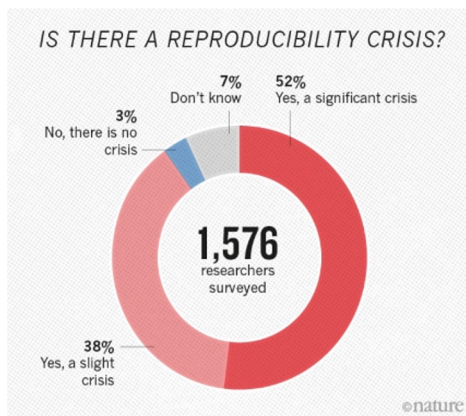
# The data problem (improper data handling)

## Risks

- Information is lacking to reproduce studies
- If studies are reproduced, errors appear, often *after* publication and clinical trials based on the outcomes
- Errors concern data documentation and mislabeling, often a lack of information to do auditing/quality assurance at all
- With potential wrong interpretation and conclusions, even dangerous testing on patients

## Conclusions

- Finding errors is time costly
- We need a culture shift in how to address scientific mistakes
- Nothing more human than making mistakes ... but make sure to facilitate reproducing your analysis



## Reproducibility crisis

- >> Selective reporting
  - >> Pressure to publish
  - >> Insufficient oversight and mentoring
  - >> Supporting data, methods and code not available
- Scientific crisis ... Who is going to believe researchers?*

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



# The data problem (importance of FAIR metadata)

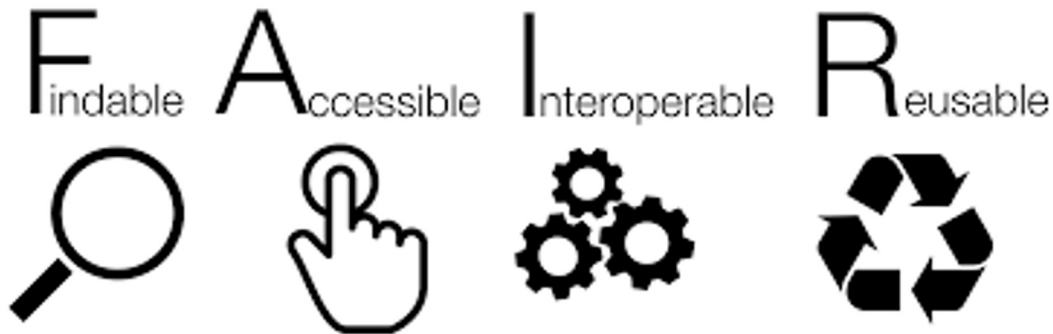
## How many ways can you say 'female'?

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 female	femlale
diploid female	female(gynococious)	remale	metafemale
f	femele	semi-engorged female	sterile female
famale	female, pooled	sexual oviparous female	normal female
femal	femalen	sterile female worker	sf
female	females	strictly female	vitellogenic replete female
female - worker	females only	tetraploid female	worker
female (alate sexual)	gynococious	thelytoky	hexaploid female
female (calf)	healthy female	female (gynococious)	female (f-o)
hen	probably female (based on morphology)		

female (note: this sample was originally provided as a \"male\" sample to us and therefore labeled this way in the brawand et al. paper and original geo submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a female individual)",



# FAIR data stewardship as a solution



## Reproducibility crisis

- >> Selective reporting
- >> Pressure to publish
- >> **Insufficient oversight and mentoring**
- >> **Supporting data, methods and code not available**

A data steward can help you ... even after you have left the organisation



## Ten practices: 'prepare well to prevent data disaster'

### Ten Practices

Keep Sufficient Documentation

Organize Files and Name the Consistently

Version the Files

Create a Security Plan

Define Roles and Responsibilities

Back up the Data

Identify Tool Constraints

Close Out the Project

Put the Data in a Repository

Write These Conventions Down in a DMP

Practical article about [Foundational Practices of Research Data Management](#)

**Exercise (30 minutes):** work in pairs and use the [shared notes](#) to

- reflect on what you are already doing
- brainstorm about what you could improve

... followed by a [joint discussion \(15 minutes\)](#)





## Practice 1. Keep sufficient documentation

### What may be improved in your current practices?

- Create documentation standards/templates to ensure recording of the same information
- Take the 'outsider' perspective: can others understand my data activities?
- Document so that research is reproducible (document more and more)
- Don't wait: document your data right away



## Practice 2. Organise files and name them consistently

### What may be improved in your current practices?

- Record the structure you choose, and create standards/templates to ensure organising in the same way
- Separate raw, analysed, processed data
- Separate ongoing from closed work (milestone versions)
- If you use abbreviations, make sure to explain them
- Keeping folder and file names as short as possible
- Apply the same to physical data (samples)
- Don't wait: organise your data right away





## Practice 3. Version the files

### What may be improved in your current practices?

- Use a version control system such as Git
- Alternatively, table at the beginning of a document with version, changes done and who (in parallel to naming the versions)
- Having a common group-platform for corrections or annotations
- Especially when working with others: add date to version (to prevent all those multiple 'final' versions)
- More consistent version names, standardise versioning



## Practice 4. Create a security plan

### What may be improved in your current practices?

- Regularly review the security plan, together with the institution's security officer
- Be aware of your own responsibilities in the plan
- Explore your institution's storage solutions
- Discuss what happens with the data after your leave
- Make sure not to lose passwords, encryption keys or key files



## Practice 5. Define roles and responsibilities

### What may be improved in your current practices?

- Formally appoint roles and responsibilities
- Document responsibilities
- Standardise procedures (SOPs) before the project
- Document change management during the project
- Define good data management practices (use these workshop tips)
- Knowledge exchange and training (mentors)
- Draft out 'best practices' (for instance for onboarding and exit processes)
- Realise you have multiple roles: as project member, as institute member, etc.



## Practice 6. Back up the data

### What may be improved in your current practices?

- Commit to the (automated) backup procedures of your institution
- Periodically confirm that backups are functioning properly to avoid unpleasant surprises when trying to recover data in a crisis
- Learn how to restore files (condition: good data organisation, including clear file names etc.)
- Implement the 3:2:1 rule: have at least three copies of your data; store the copies on two different media; keep it safe with one backup copy offsite
- Back up on laptop: don't lose your laptop



## Practice 7. Identify tool constraints

### What may be improved in your current practices?

- Evaluating in the security plan both individual software tools and any potential security gaps between tools
- Also include these discussions in the documentation of the project
- Better organising of the data once collected
- Find Python packages or R packages for conversion or data visualisation
- To develop a secure and feasible research workflow
- Discussing a more rapid technology for data transfer



## Practice 8. Close out the project

### What may be improved in your current practices?

- Storage hardware should be updated before the hardware format falls out of regular usage and becomes difficult to read
- Regularly structure and organise outputs
- Make a master copy or specific copies with raw data and final data utilised for a specific project
- Implement the snapshotting at key points of the project



## Practice 9. Put the data in a repository

### What may be improved in your current practices?

- Discuss with group to using external repositories, once applicable
- Make a plan when to put data in a repository
- Decide about the selection of data
- Training about data management practices for data archiving would be helpful
- Make data publically available for reproducibility and reuse
- Checking the understandability of final data
- Discussing about the responsible person (i.e. funder, PI, ESR) for general data sharing
- Check whether the tool operator can provide access to the data to others at final stage of the project



## Practice 10. Write these conventions down in a DMP

### What may be improved in your current practices?

- Creating a DMP if you didn't do so yet, per project or as joint effort
- For shared projects discuss DMP setup with others involved
- Have your DMP checked by your local RDM support office
- Periodically discuss and update the DMP with the group and with your PI





# Introduction to ELIXIR RDM resources

**RDMkit**

Data management

About

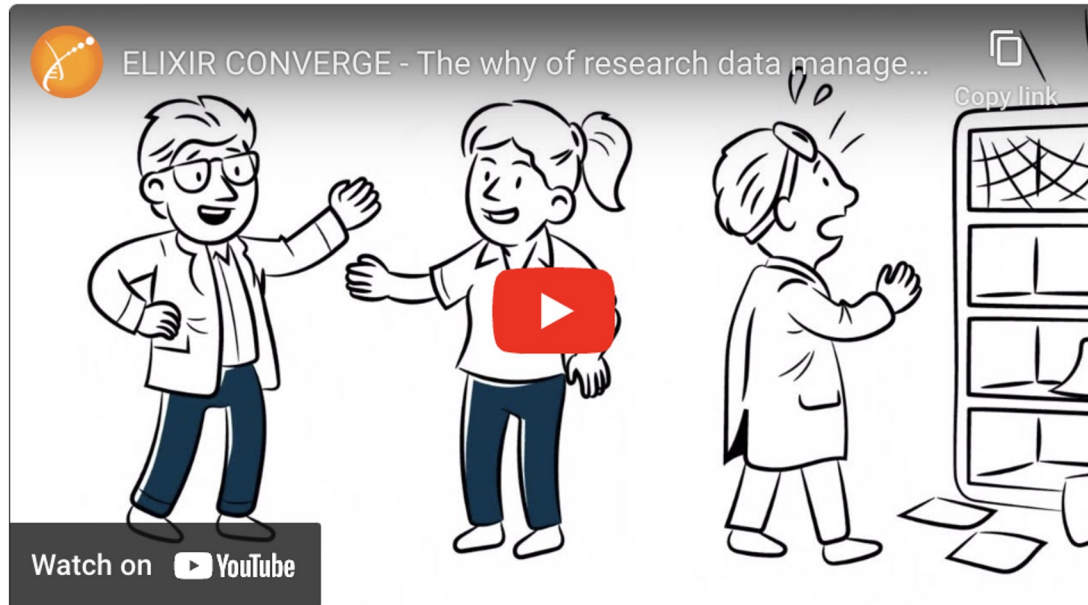
Contribute

 GitHub

The Research Data Management toolkit for Life Sciences

Best practices and guidelines to help you make your data FAIR (Findable, Accessible, Interoperable and Reusable)

## Research Data Management at glance



# Introduction to ELIXIR RDM resources



## Data life cycle

Start here to get an overview of research data management based on stages in the data life cycle.



## Your role

Identify your role in research data management, find data management resources relevant for you, and information to help you progress in your career path.



## Your domain

Learn about data management tasks that affect your domain or research community, and the solutions adopted to address them.



## Your tasks

Find guidelines and solutions for tackling common data management tasks.



## Tool assembly

Find concrete combinations of tools and resources assembled into an ecosystem for research data management.



## National resources

Find pointers to country specific information resources and national research data management practices.



## All tools and resources

Browse the RDMkit's catalogue of tools and resources for research data management.



## All training resources

Browse all training resources mentioned in RDMkit pages.



# Introduction to ELIXIR RDM resources

Your tasks

## Ethical aspects

Ethics refers to moral principles and norms that help us identify right from wrong within a particular context. Ethical issues/concerns typically arise when these principles conflict. Navigating through research involving human participants, such ethical concerns may appear when accessing a sensitive nature, for example health or personal data. Ethics, however, goes beyond with legal obligations, and the collection and use of data.

The [Open Data Institute](#) narrows 'ethics' in the RDM context to:

"A branch of ethics that evaluates data practices with the potential to adversely impact society – in data collection, sharing and use."

## Which aspects of RDM might raise ethical concerns?

### Description

Ethical issues refer to moral principles and standards that guide human conduct and define right or wrong within a particular context.

### Considerations

- There are different aspects in the management of research data that can raise ethical concerns to distinguish between ethical issues and legal behaviour.
  - Ethical standards may vary across cultures, disciplines, and professional organisations expected to adhere to these ethical principles even if certain practices are not enshrined in law. Often these standards are collected in declarations and guidelines, which are not legally binding.
  - Legal behaviour, on the other hand, refers to compliance with applicable laws, and legal requirements provide a baseline level of conduct that researchers must not breach. However, legal compliance does not necessarily guarantee ethical behaviour may be legally permissible but raise ethical concerns, while others may be explicitly prohibited by specific legislation.

Your tasks

## Data organisation

## What is the best way to name a file?

### Description

Brief and descriptive file names are important in keeping your data files organised. A file name is the principal identifier for a file and a good name gives information what the file contains and helps in sorting them, but only if you have been consistent with the naming.

### Considerations

- Best practice is to develop a file naming convention with elements that are important to your project already when the project starts.
- When working in collaboration with others, it is important to follow the same file naming convention.

### Solutions

#### Tips for naming files

- Balance with the amount of elements: too many makes it difficult to understand vs too few makes it general.
- Order the elements from general to specific.
- Use meaningful abbreviations.
- Use underscore (\_), hyphen (-) or capitalized letters to separate elements in the name. Don't use spaces or special characters: ?!&, \* % # ; \* ( ) @\$ ^ ~ ' { } [ ] < >.



# Introduction to ELIXIR RDM resources



## Links to FAIR Cookbook

### FAIRCOOKBOOK

FAIR Cookbook is an online, open and live resource for the Life Sciences with recipes that help you to make and keep data Findable, Accessible, Interoperable and Reusable; in one word FAIR.

Creating a data/variable dictionary

Creating a metadata profile

Surveying extraction, transformation, load (ETL) tools

## Links to DSW




With [Data Stewardship Wizard](#) (DSW), you can create, plan, collaborate, and bring your data management plans to life with a tool trusted by thousands of people worldwide — from data management pioneers, to international research institutes.

How will you do file naming and file organization?

Are you using a filesystem with files and folders?

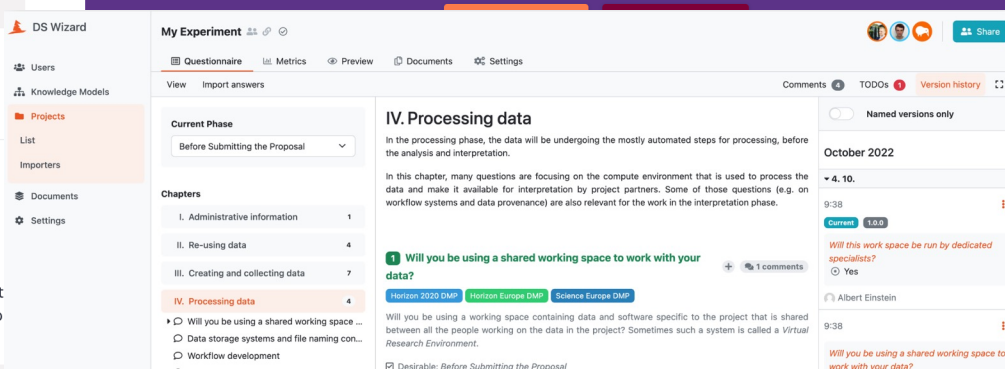
Data storage systems and file naming conventions

## Tools and resources on this page

Tool or resource	Description	Related pages	Registry
<a href="#">Bitbucket</a>	Git based code hosting and collaboration tool, built for teams.		
<a href="#">Bulk Rename Utility</a>	File renaming software for Windows		

## The FAIR Cookbook for FAIR doers

An online, **open** and **live** resource for the Life Sciences with recipes that help you to make and keep data Findable, Accessible, Interoperable and Reusable; in one word FAIR.



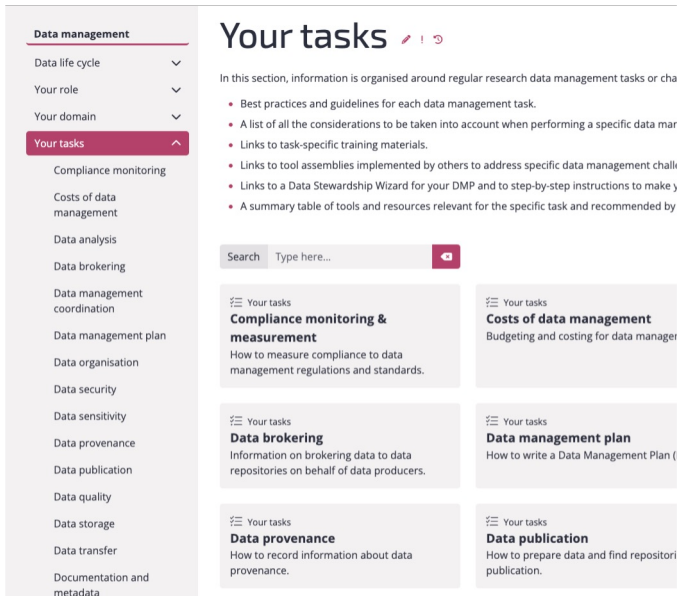
The screenshot shows a web interface for 'My Experiment'. It features a sidebar with navigation options like 'Users', 'Knowledge Models', 'Projects', 'List', 'Importers', 'Documents', and 'Settings'. The main content area is titled 'IV. Processing data' and includes a 'Current Phase' dropdown menu set to 'Before Submitting the Proposal'. Below this is a 'Chapters' list with items like 'I. Administrative information', 'II. Re-using data', 'III. Creating and collecting data', and 'IV. Processing data'. A specific question is highlighted: 'Will you be using a shared working space to work with your data?'. The interface also shows a 'Version history' section for 'October 2022' with a 'Current' version 1.0.0 and a comment from Albert Einstein.



standards, databases, policies



## RDMkit: best practices & guidelines to help you make your data FAIR



**Data management**

- Data life cycle
- Your role
- Your domain
- Your tasks**
- Compliance monitoring
- Costs of data management
- Data analysis
- Data brokering
- Data management coordination
- Data management plan
- Data organisation
- Data security
- Data sensitivity
- Data provenance
- Data publication
- Data quality
- Data storage
- Data transfer
- Documentation and metadata

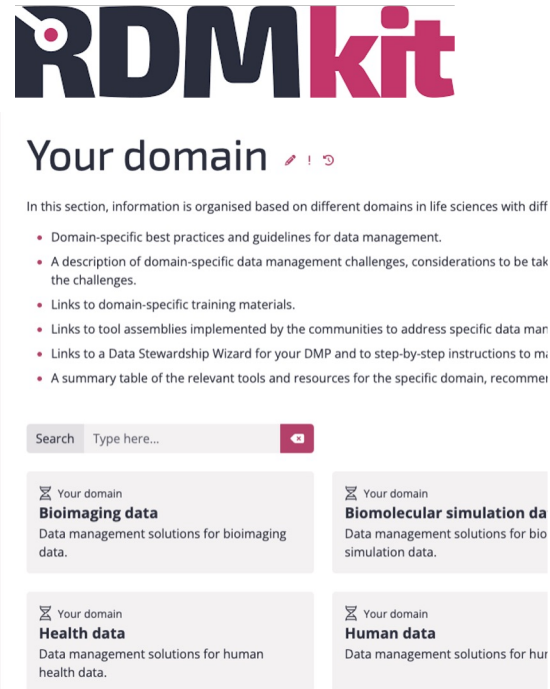
### Your tasks

In this section, information is organised around regular research data management tasks or challenges.

- Best practices and guidelines for each data management task.
- A list of all the considerations to be taken into account when performing a specific data management task.
- Links to task-specific training materials.
- Links to tool assemblies implemented by others to address specific data management challenges.
- Links to a Data Stewardship Wizard for your DMP and to step-by-step instructions to make your DMP.
- A summary table of tools and resources relevant for the specific task and recommended by the community.

Search Type here...

- Compliance monitoring & measurement**  
How to measure compliance to data management regulations and standards.
- Costs of data management**  
Budgeting and costing for data management.
- Data brokering**  
Information on brokering data to data repositories on behalf of data producers.
- Data management plan**  
How to write a Data Management Plan (DMP).
- Data provenance**  
How to record information about data provenance.
- Data publication**  
How to prepare data and find repository publication.



# RDMkit

### Your domain

In this section, information is organised based on different domains in life sciences with different challenges.

- Domain-specific best practices and guidelines for data management.
- A description of domain-specific data management challenges, considerations to be taken into account, and recommended tools and resources.
- Links to domain-specific training materials.
- Links to tool assemblies implemented by the communities to address specific data management challenges.
- Links to a Data Stewardship Wizard for your DMP and to step-by-step instructions to make your DMP.
- A summary table of the relevant tools and resources for the specific domain, recommended by the community.

Search Type here...

- Bioimaging data**  
Data management solutions for bioimaging data.
- Biomolecular simulation data**  
Data management solutions for biomolecular simulation data.
- Health data**  
Data management solutions for human health data.
- Human data**  
Data management solutions for human data.

RDMkit: data management [tasks](#) and [domains](#)

**Exercise (25 minutes):** work in pairs and use the [shared notes](#) to

- reflect on what you will be doing differently after today
- brainstorm about the activities, facilities or services the NUTRIOME team could organize or offer to help you change things

... followed by a [joint discussion \(15 minutes\)](#)



## For your own use after the workshop: Summary slides (per topic of the data life cycle)

Special thanks to

- Esther Plomp, Santosh Ilamparuthi & Marta Teperek from TU Delft
- Among others, Cees Hof (DANS), Jasmin Böhmer (UMCU), Christine Staiger (WUR) from the Helix DAS data stewardship training program
- Inge Slouwerhof, Maaïke Messelink & Tina Reiling from Radboud University
- Mateusz Kuzak, Netherlands eScience Center

for reusing their slides





## Data life cycle: planning data



Design research; plan data management; plan consent for sharing; plan data collection, processing protocols and templates; explore existing data resources

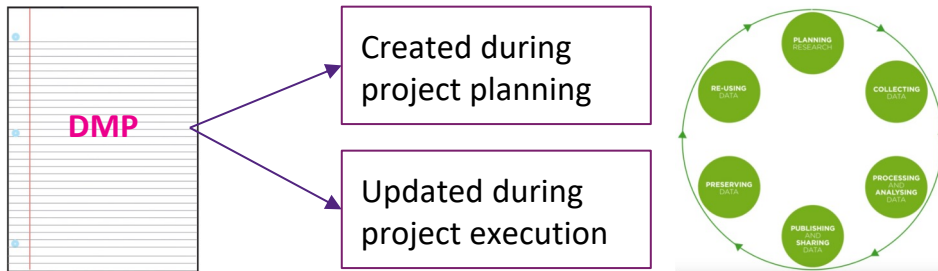
### Topics

- Writing a data management plan
- Informed consent procedures



## Summary slide: writing a data management plan

A data management plan (DMP) is a document which outlines how research data will be managed over the course of a research project



Organisation; roles; costs; existing data; collection process; informed consent; ethics committee; privacy; security; data types; tools for storing, sharing; structuring data; metadata and documentation; preserving; giving access to data etc.

Nice to watch: the [data management planning knowledge clip by Oncode, including extra materials](#)

### Benefits (vs. 'ticking boxes')

- Makes research more efficient; think and decide timely about RDM issues
- Use it as a dynamic document; use it as a discussion document
- Useful in meetings for monitoring progress of your research

### Policies & DMPs

- DMPs are used to check the 'local feasibility' of a study, i.e. a quality assurance (safety, quality, expertise) of trials executed, to acquire approval from the Executive Board
- Most funders require researchers to submit a DMP in the context of a grant application, in addition to a Data Management Paragraph





# 10 TOPICS IN A *good* DATA MANAGEMENT PLAN *template*

- 1 PRE-PROJECT ACTIVITIES**  
Questions about necessary pre-project activities: planning, budgeting, ethical & legal constraints, responsibilities.
- 2 COSTS**  
Questions on data management related costs, e.g. costs of data storage, transcription, anonymisation, archiving, training et cetera.
- 3 DATA COLLECTION/REUSE**  
Questions on all aspects of data gathering: types and sources of data, volume, file formats, quality assurance, sustainability, reproducibility, required software.
- 4 DATA DOCUMENTATION**  
Questions on the description of data, metadata standards, interoperability, required for FAIR data.
- 5 STORAGE & ORGANISATION**  
Questions on storage facilities, backups, file structure, file names, versioning.

- 6 ACCESS TO DATA**  
Questions on what data will be accessible during the project to which project team member, and about the prevention of unauthorised access.
- 7 SHARING & PUBLICATION**  
Questions on sharing of data, if any, with others during and after the project, reuse potential, findability.
- 8 ARCHIVING OF DATA**  
Questions on the where and how of data archiving, retention periods, licences, persistent identifiers.
- 9 ETHICAL & LEGAL ASPECTS**  
Questions on ethical and legal aspects of data collection, storage, sharing, publication and archiving - including privacy and intellectual property rights.
- 10 GENERAL INFORMATION**  
Questions about the research project, updates and versioning of the data management plan.

A DMP template should provide guidance on the questions, e.g. references or examples, and contact information of (local) research data management support.



<https://lcrdm.nl/data-stewardship/>, bottom of page





## 10 TIPS FOR WRITING A DATA MANAGEMENT PLAN

1

### START EARLY

Read the guidance and ask for advice early on in the process, as writing a DMP may take some time

2

### CONSIDER REUSE

Think about reusing existing data. Describe what you will need to know about your data five years from now

3

### CHECK POLICIES

Talk to your supervisor or lab members about existing data management policies and standards

4

### MAKE USE OF SUPPORT

Use your in-house support services like RDM Support, the Library, IT department or legal desk

5

### THINK BROAD

Also address software code, algorithms and any other valuable research assets in your DMP

6

### COPY WHERE YOU CAN

Look at other (submitted) plans and copy when appropriate

7

### BE UNIQUE WHERE NEEDED

Since every research project is unique, so are the data it generates. Copying from sample DMPs is not sufficient

8

### BE CONCRETE

Make your answers as concrete as possible. Show that you have consulted RDM experts

9

### SAY SO IF YOU DON'T KNOW

Indicate what you do not yet know and how you will resolve these questions later

10

### UPDATE

DMPs add to the planning of your research methods. Therefore define, carry out and update your DMP just as you would any method



Data Horror Week 2020



## Learn about data management concepts in this online Data Horror Escape Room

📅 October 30, 2020 • 📁 Projects • ⌚ 6 min read

Our new digital escape room introduces some basic data management concepts and offers a fun way to educate and to prompt discussions with researchers.

**Guest co-authors: Lena Karvovskaya and Elisa Rodenburg, VU Amsterdam, with Joanne Yeomans, Leiden University Libraries.**

The [Data Horror Escape Room](#) was a collaboration between Vrije Universiteit Amsterdam (Lena Karvovskaya and Elisa Rodenburg), Leiden University Libraries (Joanne Yeomans), and Eindhoven University of Technology (Anne Aarts and Bart Aben).

[Learn about data management concepts in this online Data Horror Escape Room - Digital Scholarship Leiden](#)



## Tool: DMP online

2

# DMPonline

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

 DMP  
ONLINE



# Tool: data stewardship wizard

Yet another DMP tool? NO! This is a wizard!

“A help feature of a software package that automates complex tasks by asking the user a series of easy-to-answer questions”

... Just like a checklist helps an airplane pilot (the expert!) to make sure he doesn't forget anything!


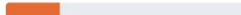



- *Irritating*: knowing that there are tools/solutions for FAIR but you can't find them
- *Painful*: not knowing that the problem you experience is already solved by others
- *Dangerous*: not realising the problems you might run into

The DSW is the “online data steward” (beyond office hours) that helps researchers to ask the necessary questions and find answers

- As an expert, not as a bureaucratic, static (DMP) template (but it *will* generate those templates if/as you need them!), based on the FAIR principles

## FAIR Metrics

You can **evaluate answers** in each questionnaire to get an overview of how good you are doing in terms of **FAIR metrics**. Thus you are able to reconsider your decisions already in the planning phase to make your research more FAIR.

Metric	Measure
Findability	0.70 
Accessibility	0.22 
Reusability	1.00 
Good DMP Practice	0.67 
Openness	0.69 

1.a.2 Is this a standard data format used by others in this field? + !

Desirable: Before Submitting the DMP

a. No ☰

**Interoperability**

b. Yes

**Interoperability**





DS Wizard

- Knowledge Models
- Questionnaires
- KM Editor
- Storage Costs Evaluator

---

- Help >
- Nikola Tesla >

« Collapse sidebar

My Experiment (Common DSW Knowledge Model, 1.4.0) Save

**Current Phase**  
Before Submitting the Proposal

## I. Design of experiment

Before you decide to embark on any new study, it is nowadays good practice to consider all options to keep the data generation part of your study as limited as possible. It is not because we can generate massive amounts of data that we always need to do so. Creating data with public money is bringing with it the responsibility to treat those data well and (if potentially useful) make them available for re-use by others.

**Chapters**

- I. Design of experiment** 6
- II. Data design and planning 5
- III. Data Capture/Measurement 3
- IV. Data processing and curation 1
- V. Data integration 1
- VI. Data interpretation 1
- VII. Information and insight 8

**More**  
Summary Report

**1 Is there any pre-existing data?** + !

Are there any data sets available in the world that are relevant to your planned research?

Desirable: Before Submitting the Proposal  
Data Stewardship for Open Science: atq

a. No

b. Yes

Clear answer

**1.b.1 Will you be using any pre-existing data (including other people's data)?** + !

Will you be referring to any earlier measured data, reference data, or data that should be mined from existing literature? Your own data as well as data from others?

Desirable: Before Submitting the Proposal  
Data Stewardship for Open Science: gzi

a. No

b. Yes

Clear answer

**1.b.1.b.1 What reference data will you use?** + !

Much of today's data is used in comparison with reference data. A genome for instance is compared with a reference genome to identify genomic variants. If you use reference data, there are several other issues that you should consider. What are the reference data sets that you will use?

**Current Phase**

- Before Submitting the Proposal
- Before Submitting the DMP
- Before Finishing the Project

**Chapters**

- I. Administrative details 2
- II. Re-using data 3
- III. Creating and collecting data 8**
- IV. Processing data 3
- V. Interpreting data 1
- VI. Preserving data 6
- VII. Giving access to data 3

**More**

- Summary Report



## Guidance

Our smart questionnaires will effortlessly guide you through the vast knowledge of data stewardship by asking you relevant **questions**, offering **hints**, **multimedia contents**, **external resources** and **community help**.


## Data Stewardship Hints

With kind permission of **Taylor & Francis Group**, we provide information and hints from **Data Stewardship for Open Science** by **Barend Mons** directly in the Data Stewardship Wizard.

<https://ds-wizard.org/>

### 1 Is there any pre-existing data?

a. No

b. Yes 

 Clear answer

You know that this is very unlikely? This question is not only about yourself, but also reference data or data that should be mined from refer to related data, e.g. other databases where you usually "quickly" integrated, especially if you need to do such lookups multiple times!



Data Stewardship for Open Science: Chapter 1.1

With kind permission of  
 CRC Press  
Taylor & Francis Group

### Is there pre-existing data?

#### What's up?

For many decades if not centuries, virtually every experiment started with the collection or creation of 'observations' and in fact data. In social sciences and humanities the tendency to 'reuse' data that had been created earlier, in all kinds of surveys and increasingly of course from sources such as social media maybe already somewhat more established. However, in many of the hard experimental sciences, the generation of new data specifically generated to answer a hypothetical question is still so commonplace that careful thinking about the actual need to generate new data may just not be on the radar screen. Obviously, data creation will need to continue, but increasingly we have to ask the question whether such new data are absolutely necessary to answer the question we want to answer. With more and more data becoming available in reusable format, there may well be existing data collections, 'Other People's' Data and associated Services (OPEDAS) that without or with some extra effort needed, can answer at least part of the question or least may be crucial for the interpretation of your own data.

#### Do

- Search for data sets (OPEDAS) that may be re-usable and can help you to reduce the number of new data sets you may have to generate (and steward later on).
- Include annotated collections of data and curated databases in your search.
- Check the accessibility and license situation attached to the relevant data sets you found.
- Check their interoperability. They may be relevant but not interoperable with your analysis pipelines. In that case you may have to extract, transform and load (ETL) them or decide that –although relevant– they are not reusable for your purpose.
- Ensure that using OPEDAS will not restrict in any way the use of your results later on, including copyright and freedom to operate on the request of IPR.
- Check how to cite and acknowledge OPEDAS.
- Consider to actively involve OPEDAS owners in your research in order to make optimal use of their data.
- Speak to colleagues who did similar experiments before to find out about potential OPEDAS you may consider to use.

# Tool: 23 Things (and more)



National Coordination Point  
Research Data Management



## 23 Things for Researchers and PhD Candidates

An overview of practical resources and tools that you can begin using today to incorporate research data management into your research workflows.

### Contents

Research Data Management

Data Management Plans

Personal & Sensitive Data

Metadata & Data Documentation

Digital Preservation & Data Repositories

Data Licensing

Citing Data

Community of Practice

Learning Resources

... to help researchers and PhD candidates engage in research data management!

Nice to browse: [the 23 Things for researchers and PhD candidates](#) and the [23 Things tool](#)





# Summary slide: informed consent procedures

Agreement between the researcher and the data subject (i.e. participant), including:

- The data subject is **informed**: provide information that is received and understood (information brochure)
- The data subject gives **consent**: you need an explicit statement that the data subject freely agrees to participation in the research project

## Legal perspective

As a legal base for collecting personal data (GDPR)



## Ethical perspective

Participant is informed and thus enabled to make a voluntary decision about accepting or declining participation in research



Written informed consent (often: obliged for WMO research)

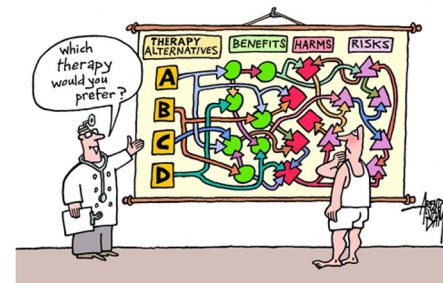


Oral informed consent



Online informed consent

**Always check the informed consent templates from your institute/ethical committee**



*informed consent*



# Informed consent procedures

## What should an informed consent procedure contain?

- The name and contact details of the researcher and data protection officer
- Research procedures
- Benefits, discomfort and risks
- The purposes of the processing of the personal data
- The (categories of) recipients of the personal data
- The retention period for the personal data
- Data subject's rights with regard to the processing of personal data
- The right to withdraw consent
- The right to withdraw data (for example up to 2 weeks after collection)
- The right to lodge a complaint with the Data Protection Authority
- If applicable: compensation and insurance

Always check the informed consent templates from your institute/ethical committee



Written informed consent (often: obliged for WMO research)



Oral informed consent



Online informed consent



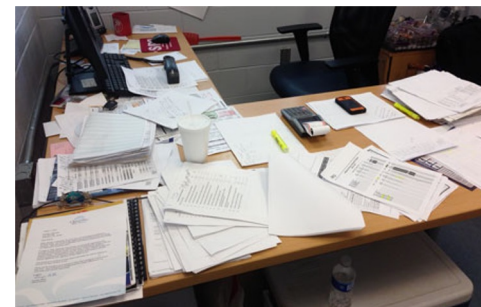
# Informed consent procedures

## Written informed consent: practical tips

- Informed consent forms contain personal data. Safe storage is important
- Physical documents: store in locked cabinet with limited people having access to the key. You have to make sure the change of unauthorised access is minimised
- Digitisation: informed consent forms can be digitised in accordance with the guidelines from your institution

The digital version should then be treated as all other digital personal data. The paper version can then be destroyed after (check requirements and time periods of your institution)

However: an exception to this are consent forms for WMO obligated research  
These may not be substituted: if they are paper, they must remain paper



# Informed consent procedures

## Online informed consent: practical tips

Keep a record of the oral consent procedure:

- Who consented (the name of the data subject or other identifier)
- When the data subject consented (a note of the time and date)
- What the data subject was told at that time (keep a copy of the presented information)
- How the data subject consented (keep a short note of the conversation at that time, describing the reply of the data subject)



# Informed consent procedures

## Oral informed consent: practical tips

Informed consent via for example an online survey such as Qualtrix or Castor

- Keep a copy (for example a PDF export) of the online completed survey that includes a time stamp
- Keep a copy of the information that was presented to the data subject at that time
- “I have read the provided information and I agree to the use of my data for the research purposes as described in the provided information”
  
- Pre-ticked boxes: the data subject has to actively consent
- Opt-out boxes (“If you don’t consent, tick this box”)
- Assume that submitting an online survey by a data subject demonstrates valid consent



DOs



DON'Ts



# Informed consent procedures

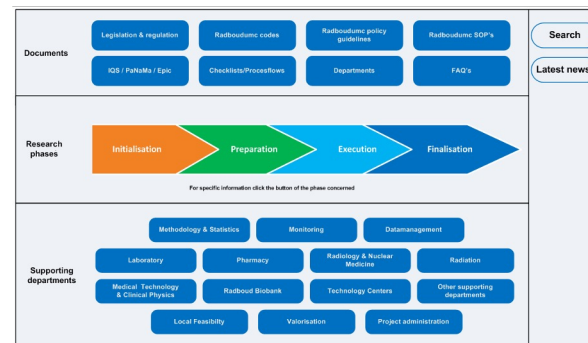
## Example from a University Medical Centre (Radboudumc)

- SOP (Standard Operating Procedures) Obtaining informed consent scientific human-based research
- SOP Recruitment and coding of study participants scientific human-based research
- [CCMO](#) template Subject Information (PIF) & consent form (obligatory for WMO research)
- Templates for pediatric research

## *Informed consent procedure (WMO obligated research)*

- Use only written informed consent forms ('wet ink signature'), no oral or online consent
- Store the original paper informed consent forms
- Store informed consent forms 20 yrs for drug trials and 15 yrs for other human related research studies

Nice to watch: the [GDPR and informed consent](#) knowledge clip by Oncode




 Central Committee on Research Involving Human Subjects

Home > Investigators > Standard research file >

### E. Information research subjects

- > E1/E2. Information leaflet and consent form research subjects
- > E3. Promotional materials research subjects
- > E4. Other informational materials
- > E5. Newsletters or letters with study results



# Informed consent procedures

## Questions to ask yourself

- Will you use an informed consent procedure in your research? Why (not)  
Reflect on the content of the informed consent form and the accompanying information document
- Did you consider data in your informed consent procedures?  
Explicitly consider long-term storage and sharing of data during and after research.
- Getting approval of an ethics committee might be a part of your research project, for instance to safeguard the rights of participants. Why (not)
- Imagine your colleague is your respondent and you are explaining the informed consent procedure of your research to him/her  
What would be your main message (“elevator pitch”)?  
Provide feedback: what would you like to hear as a respondent?



## Data life cycle: collecting data



Collect data; capture data with metadata; acquire existing third party data

### Topics

- Finding data
- Collecting data
- Preregistration
- Data security
- Personal data

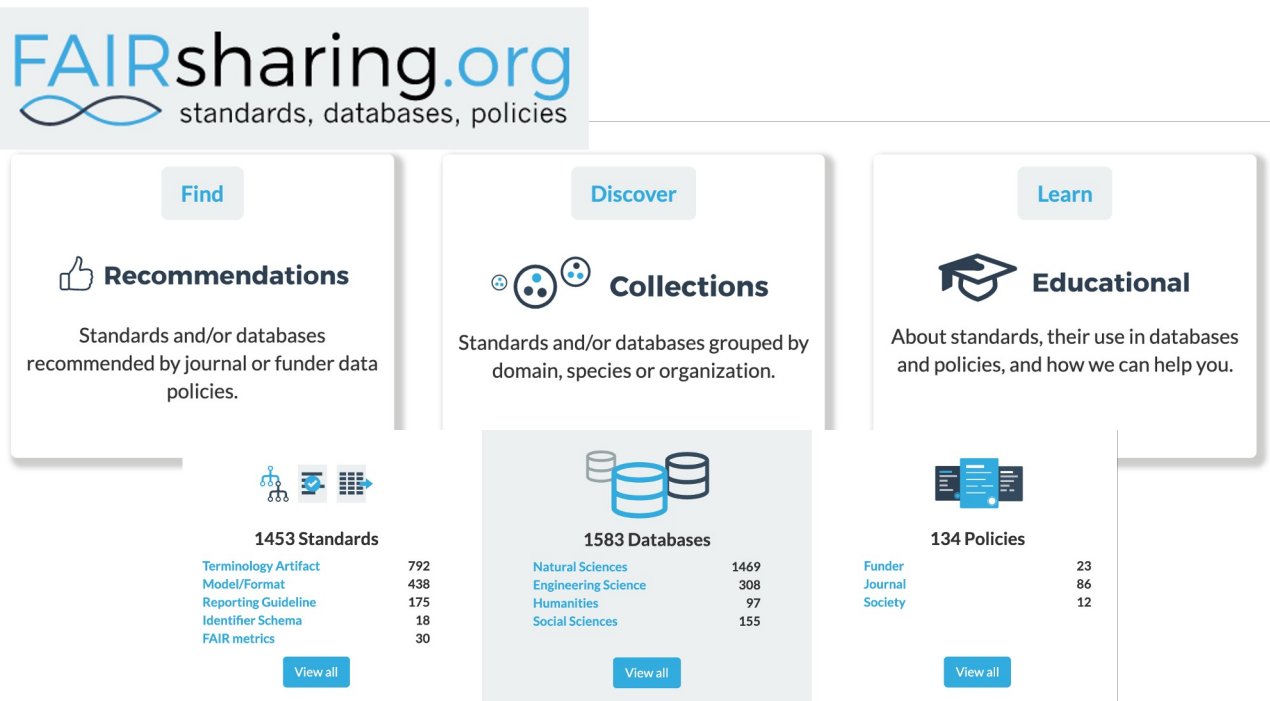




## Summary slide: finding data

If you intend to reuse existing data instead of collecting it yourself, there are good sources for potentially relevant existing data. The following (directories of) archives may be relevant sources for finding data.

- [Open Access Directory: Data repositories](#)
- [Re3data.org](#)
- [Recommended Data Repositories from Nature](#)
- [FAIRsharing](#)



**FAIRsharing.org**  
standards, databases, policies

**Find**

**Recommendations**

Standards and/or databases recommended by journal or funder data policies.

**1453 Standards**

Terminology Artifact Model/Format	792
Reporting Guideline	438
Identifier Schema	175
FAIR metrics	97
	18
	30

[View all](#)

**Discover**

**Collections**

Standards and/or databases grouped by domain, species or organization.

**1583 Databases**

Natural Sciences	1469
Engineering Science	308
Humanities	97
Social Sciences	155

[View all](#)

**Learn**

**Educational**

About standards, their use in databases and policies, and how we can help you.

**134 Policies**

Funder Journal Society	23
	86
	12

[View all](#)



# Summary slide: collecting data

<https://libguides.vu.nl/rdm/data-collection>

## Data collection

Data collection may consist of the re-use of existing data and/or the generation of new data. You can find more specific information on the re-use of existing data on the [Finding Existing Data](#) page in this LibGuide

For data to be considered valid and reliable, data collection should occur consistently and systematically throughout the course of the research project. Data collection guidelines and established methodologies should be used to gather data. Some disciplines make use of codebooks, whereas others use protocols for data gathering. These procedures help researchers collect data according to conventional methodological steps. If a research project involves multiple partners (in a consortium) it should be clear who is responsible for the collection of what (part of the) data. Important aspects of data collection include:

- Standardisation: [codebooks](#) & [protocols](#)
- Structure / organisation of the data
- [Data quality assurance methods](#)
- [Documentation & metadata](#)
- [Storage & protection](#)

This relates to the [Reproducibility](#) of your research according to the [FAIR-data principles](#).

Nice to watch: the [collecting data](#) clip by the VU



## Summary slide: preregistration

Preregistration allows the research community to get information about upcoming research, and provide feedback on the research plans, including the protocols, methods, etc. As a result, research plans can learn from and align with other research.

- Makes your science better by increasing the credibility of your results
- Allows you to stake your claim to your ideas earlier
- It's an easy way to plan for better research
- Increase transparency
- Avoid unnecessary duplication of animal studies
- Reduce reporting bias, such as publication bias and bias induced by selective outcome reporting, p-hacking and HARKing
- Increase data sharing, by
  - allowing fellow researchers and reviewers to access information on the study design, which is often lacking in publications
  - provide a platform to share details and data of otherwise unpublished animal studies
- Create opportunities for collaborative research



[Preregistration](#)

 **protocols.io**

Making it easy to share method details **before**, **during**, and **after publication**.



## Summary slide: data security

### Location of data

- Hard-copy personal and confidential data has to be kept in a locker
- In case of personal and confidential data, the university drive is often the only location allowed
- Public (free) storage or transfer services are no option for personal or confidential data

### Encrypt your devices

- Use [Filevault](#) (Mac), [Bitlocker](#) (Windows) or the open source [VeraCrypt](#) to encrypt your device

### Encrypt your data

- If it is necessary to keep your data at an untrusted storage location or you need to transport it over a network, you can encrypt your data using for instance [VeraCrypt](#) or [Axcrypt](#)

### Use VPN to transfer data

- The best way to ensure safe transport between the university network drive and your home computer is to use the VPN connection of the university

### Wipe storage devices after use

- There are simple tools that are able to retrieve information after you deleted it. To secure your device, utilities as [Permanent Eraser](#) and [CCleaner](#) can be used to destroy the data
- To erase all data from your device you can use for instance [Dban](#)



# Summary slide: personal data

## Anonymisation

Re-identification of the anonymised data combined with any other population data is impossible

## Pseudonymisation

Personal data can no longer be attributed to a specific data subject without the use of additional information (pseudonymisation key)

Short [movie](#) about [privacy by design](#)

- Data minimisation
- Data quality
- Goal setting
- Minimisation of use
- Security measures
- Transparency
- Rights of data subjects
- Liability

Make sure to check your local RDM policy, Standard Operating Procedures (SOPs), GDPR proof templates, and contact your legal office and/or privacy officer in case of questions and if you work with third parties or third party tools

## WHAT IS PERSONAL DATA?

DEFINITION AND SCOPE UNDER THE GDPR



### ANY INFORMATION

Objective (earns 10k per year); Subjective (opinion); and, Sensitive data (gay woman).



### RELATING TO

An individual, about a particular person, impacts a specific person.



### IDENTIFIED OR IDENTIFIABLE

Direct or indirectly e.g. You know me by name, direct, you know me as "a Lawyer doing these graphics", indirect.



### NATURAL PERSON

applies ONLY to a living human being. National Law may give rules for deceased persons.



### ONLINE IDENTIFIER & LOCATION DATA

Include data provided by the electronic devices we use: mobiles, cookies identifiers, IP address, others.



### TO ONE OR MORE FACTORS

Include data that when combined with unique identifiers and other info create a profile and identify a person.





## Data privacy

# 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>



# Data privacy



## Anonymisation

Re-identification of the anonymised data combined with any other population data is impossible



## Pseudonymisation

Personal data can no longer be attributed to a specific data subject without the use of additional information (pseudonymisation key)

## WHAT IS PERSONAL DATA?

DEFINITION AND SCOPE UNDER THE GDPR



### ANY INFORMATION

Objective (earns 10k per year); Subjective (opinion); and, Sensitive data (gay woman).



### RELATING TO

An individual, about a particular person, impacts a specific person.



### IDENTIFIED OR IDENTIFIABLE

Direct or indirectly e.g. You know me by name, direct, you know me as "a Lawyer doing these graphics", indirect.



### NATURAL PERSON

applies ONLY to a living human being. National Law may give rules for deceased persons.



### ONLINE IDENTIFIER & LOCATION DATA

Include data provided by the electronic devices we use: mobiles, cookies identifiers, IP address, others.



### TO ONE OR MORE FACTORS

Include data that when combined with unique identifiers and other info create a profile and identify a person.



# Data privacy

Short [movie](#) about privacy by design

## Privacy by design

Build in privacy-enhancing measures in the design of your research. Without a good plan, opportunities for data breaches are growing

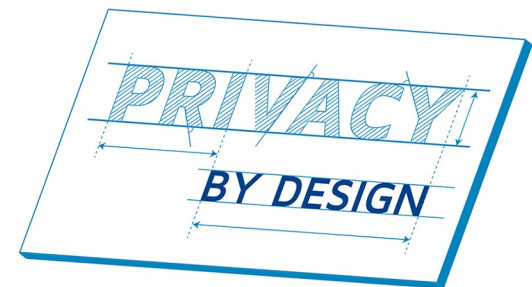
## Eight guidelines

- Data minimisation
- Data quality
- Goal setting
- Minimisation of use
- Security measures
- Transparency
- Rights of data subjects
- Liability

Make sure to check your local RDM policy, Standard Operating Procedures (SOPs), GDPR proof templates, and contact your legal office and/or privacy officer in case of questions and if you work with third parties or third party tools

## Privacy by Design

A brief instruction video with guidelines for processing personal data by RU staff (animation: Rikkert Veltman)





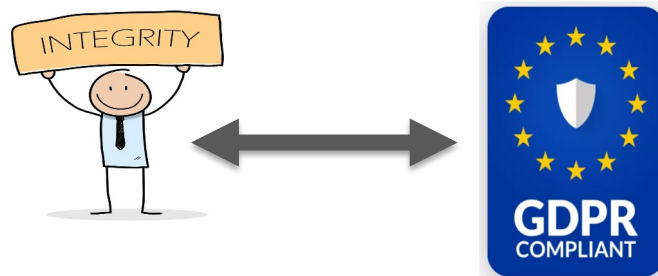
# Data privacy

## Data minimisation

Personal data in research data has to be *adequate, relevant and limited* to what is necessary for the *purposes for which they are processed*

## Review your research questionnaire and data collection process

- Are all of your questions entirely necessary to answer the research question?
- Do you actually process the data that you collect?
- Or do you collect the data because it may come in handy? (“data harking”)
- Is it possible to anonymise/pseudonymise your data?



1. Data minimisation

2. Data quality

3. Goal setting

4. Minimisation of use

5. Security Measures

6. Transparency

7. Rights of data subjects

8. Liability

# Data privacy

## Data quality

Personal data in research data has to be *of good quality, accurate and up-to-date*



You have this **dataset** containing personal data:

**QUALITY**



Participant number	Surname	Current residence	Job location
1	Jansen	Nijmegen	Utrecht
2	Velde, van der	Arnhem	Nijmegen

Participant 1 moves to Wageningen. The dataset has to be updated, since the current data is not up-to-date.

You have this **dataset** containing personal data:

**QUALITY**



Participant number	Surname	Residence at time of interview	Interview date
1	Jansen	Nijmegen	17-4-2019
2	Velde, van der	Arnhem	28-4-2019

Participant 1 moves to Wageningen. The dataset doesn't have to be updated.

1. Data minimisation
2. Data quality
3. Goal setting
4. Minimisation of use
5. Security Measures
6. Transparency
7. Rights of data subjects
8. Liability

# Data privacy

## Goal setting

In your goal setting, you must describe in detail

- *what personal data* you will be processing,
- with which *legal base*
- and for *how long* you are going to keep this data

In research:  
Informed consent



Make sure to follow your local  
organisation's registry procedures

Homework: check out the [GDPR and informed consent](#) knowledge clip by Oncode

1. Data minimisation

2. Data quality

3. Goal setting

4. Minimisation of use

5. Security Measures

6. Transparency

7. Rights of data  
subjects

8. Liability

# Data privacy

## Minimisation of use



The fewer people who have access to the personal data in your research, the better



Will a read-only access privilege be sufficient for your colleague? Do not give any editing rights



Do you have permission to share personal data? If not, you must ask for permission if you have to transfer personal data

1. Data minimisation
2. Data quality
3. Goal setting
- 4. Minimisation of use**
5. Security Measures
6. Transparency
7. Rights of data subjects
8. Liability

# Data privacy

## Security measures

Make sure that your data is well secured

When working with personal data in research, you should at least make use of *privacy protection techniques and measures*, such as:

- Encrypt devices
- Encrypt files
- Data anonymisation
- Data pseudonymisation
- Substitution of paper documents

Make sure to use your local organisation's tool and check if they are GDPR proof (via your privacy officer)



1. Data minimisation
2. Data quality
3. Goal setting
4. Minimisation of use
- 5. Security Measures**
6. Transparency
7. Rights of data subjects
8. Liability

# Data privacy

## Transparency

Transparency is a fundamental principle in the GDPR.

Transparency is about being open toward participants regarding the processing of their personal data in the research project. Transparency is an obligation of the researcher which applies to:

- Informing participants about what personal data is collected and why
- Informing participants about how they can call on their data subject rights
- Complying to data subject rights



**Informed consent procedures**

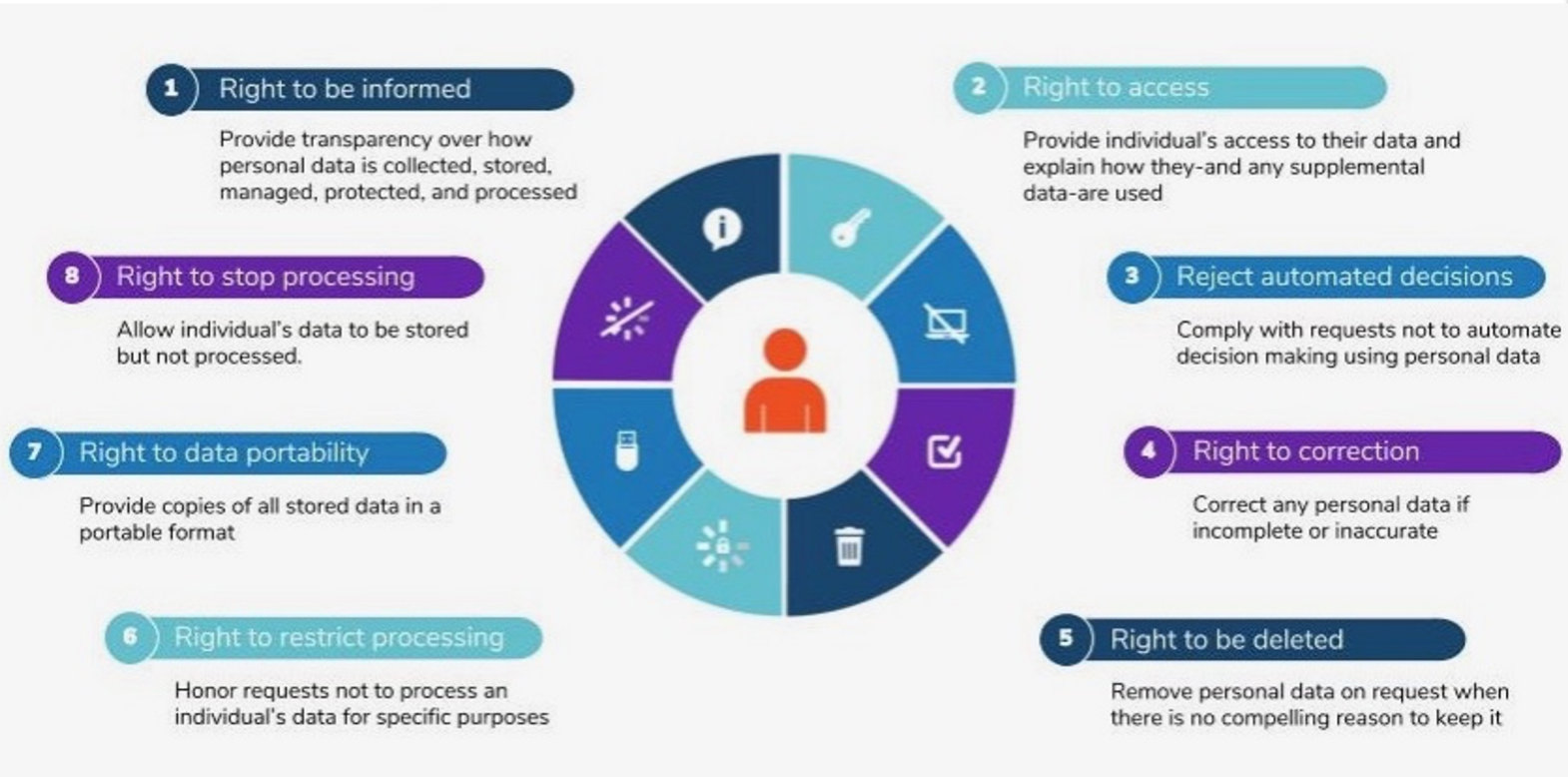


1. Data minimisation
2. Data quality
3. Goal setting
4. Minimisation of use
5. Security Measures
- 6. Transparency**
7. Rights of data subjects
8. Liability

# Data privacy

## Rights of data subjects

There are eight fundamental data subject rights



1. Data minimisation
2. Data quality
3. Goal setting
4. Minimisation of use
5. Security Measures
6. Transparency
<b>7. Rights of data subjects</b>
8. Liability

# Data privacy

## Liability

Make sure that you know who has ultimate responsibility over the collected data and how the various roles, tasks and authorizations have been established.

For that, check your local RDM policy and its corresponding responsibilities of researchers and the institute/department

1. Data minimisation
2. Data quality
3. Goal setting
4. Minimisation of use
5. Security Measures
6. Transparency
7. Rights of data subjects
8. Liability



# Data privacy



Check the National [ELSI](#) servicedesk (Dutch only)

Very useful, elaborate [example](#) webpage on the GDPR in research (Radboud University)



1. Data minimisation
2. Data quality
3. Goal setting
4. Minimisation of use
5. Security Measures
6. Transparency
7. Rights of data subjects
8. Liability

SURF has a short 20 minute e-learning module on [privacy in research](#)

## Data life cycle: processing and analysing data



Enter, digitize, transcribe and translate data; check, validate, clean, anonymize; derive data; describe; manage and store data; analyse and interpret data; produce research output; cite data sources

### Topics

- Data storage
- Data organisation
- Data versioning
- Data documentation



# Data horror stories

**Iestyn Shapey**  
@iestyn\_shapey

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.

**Sam Giles**  
@GilesPalaeoLab

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

- 19% Every day ☺
- 18% Every week
- 32% Errr every few months?
- 31% I bought a drive once...

**CASH REWARD**  
for returning my lost backpack



- Black (AK) Burton Rucksack
- Lost on Friday 15. July at 8 pm in the Pantons Arms pub 43, Pantons St. Cambridge
- Containing a laptop (white MacBook), a black external hard drive and scientific research documents

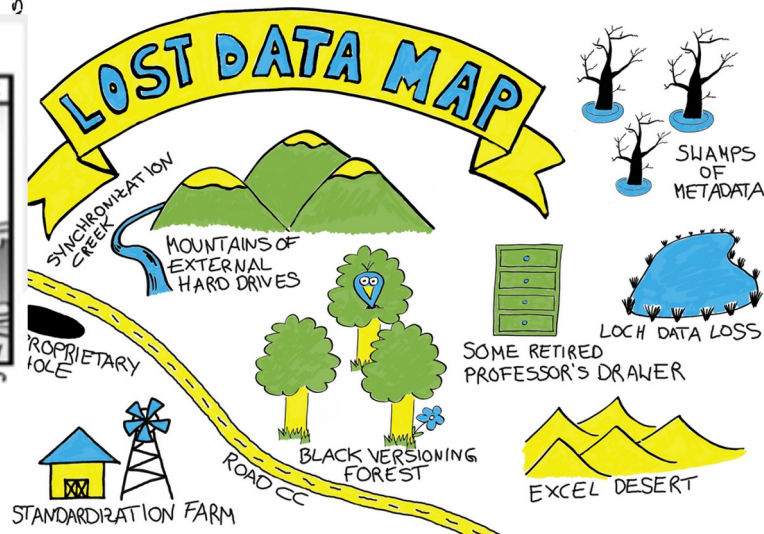
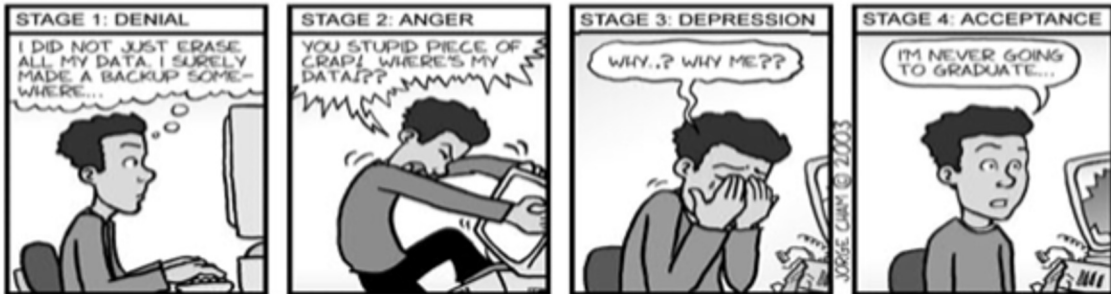
The external hard drive is VERY important to me as it contains 5 years of research data which are crucial for my PhD thesis!!!

If you found it, I would be extremely grateful if you could return it to the Pantons Arms or contact me on: 07804430054 (ar456@cam.ac.uk)

Thank you!!

## THE FOUR STAGES OF DATA LOSS

DEALING WITH ACCIDENTAL DELETION OF MONTHS OF HARD-EARNED DATA



# Data horror stories

Original slide by Marta Teperek

## 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?



7

## 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/>





## Summary slide: data storage

The ideal tools allow you to do much more than only storing ...

- Import data sources, organise data, store data securely
- Collaborate on data
- Access anywhere, any time, on any device
- Data processing and analysis tools
- Fully scalable data quantity and compute
- High-performance computing
- Controlled access and logging
- Use, manage, combine and re-use data
- Use the research tools and applications you want
- Handle multiple types of data: clinical, images, omics, etc.
- Generate virtual workspaces for researchers
- Pseudonymization software

Such as ... virtual/digital research environments, secured university storage, SURF services, or: ask the data steward wizard (<https://ds-wizard.org/>)



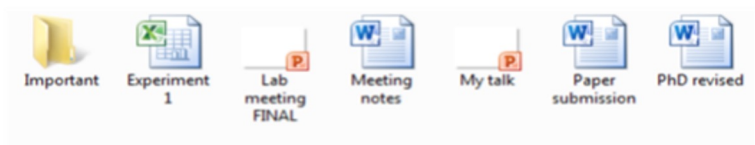
The screenshot shows the DS Wizard interface. On the left is a navigation menu with items like 'DS Wizard', 'Knowledge Model Editor', 'Knowledge Models', 'Questionnaires', 'Documents', and 'Storage Costs Evaluator'. The main content area is titled 'My Experiment' and includes a 'Current Phase' dropdown set to 'Before Submitting the Proposal'. Below this is a 'Chapters' list with items like 'I. Administrative details', 'II. Re-using data', 'III. Creating and collecting data' (highlighted), 'IV. Processing data', 'V. Interpreting data', 'VI. Preserving data', and 'VII. Giving access to data'. A 'More' section contains 'TODOs'. On the right, there is a section titled 'III. Creating and collecting data' with a sub-section '1. What data formats/types will you be using?' containing text about data formats and a 'Data format/type' field with 'RDf/XML Syntax Specification' and a 'FAIRsharing' link.

**SURFdrive: store and share your files securely in the cloud**

**Research Drive: securely and easily store and share research data**

**SURFfilesender: send large files securely and encrypted**

## Summary slide: data organisation



In 3 years time would you know what these are?



Original slide by Marta Teperek

# 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

## File naming convention:

### TILS Document Naming Convention

Document naming for the TILS Division should follow this convention:

GDL\_TILSDocNaming\_V1\_20090612.docx

A prefix shows the document type

The document title describes the content

The version number

The date in the format yyyymmdd

## Data organisation

# NO

myabstract.docx

Joe's Filenames Use Spaces and Punctuation.xlsx

figure 1.png

fig 2.png

JW7d^(2sl@deletethisandyourcareerisoverWx2\*.txt

# YES

2014-06-08\_abstract-for-sla.docx

joes-filenames-are-getting-better.xlsx

fig01\_scatterplot-talk-length-vs-interest.png

fig02\_histogram-talk-attendance.png

1986-01-28\_raw-data-from-challenger-o-rings.txt







## Data organisation

### three principles for (file) names

machine readable

<https://speakerdeck.com/jennybc/how-to-name-files>

human readable

plays well with default ordering

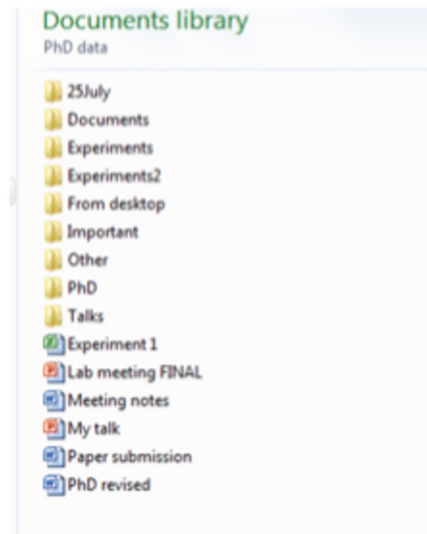
easy to implement NOW

payoffs accumulate as your skills evolve  
and projects get more complex

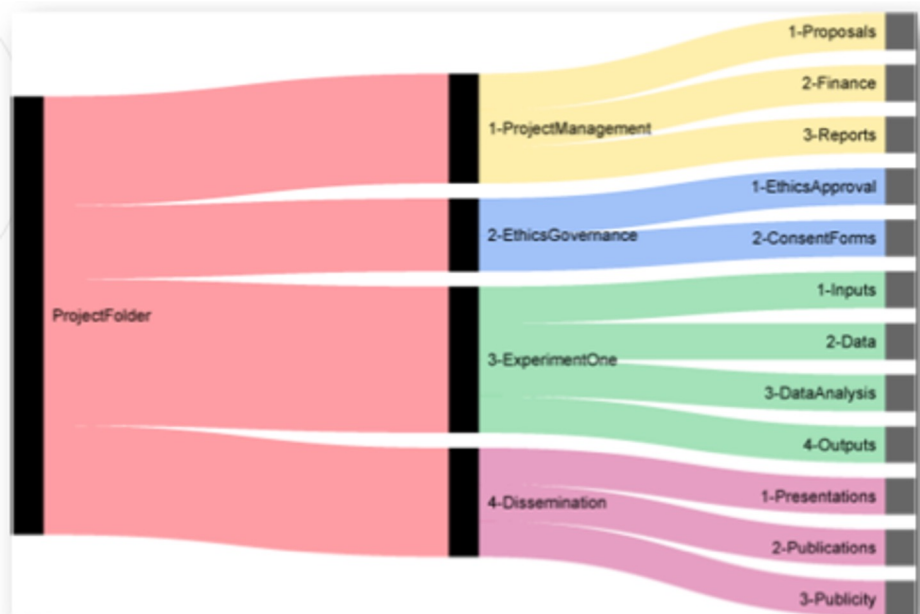
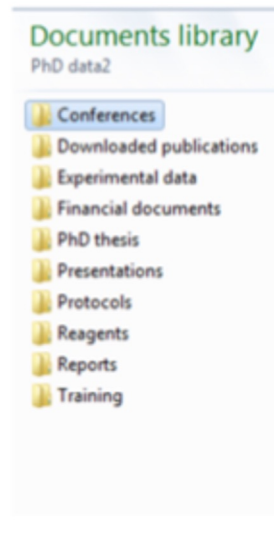


# Data organisation

Example-A



Example-B



## Summary slide: data versioning

- Use a 'revision' numbering system. Example: v03\_01
- Add information to identify the individual who has made the amendments & a date stamp. Example: 20190508\_datav01\_SJ
- Decide how many versions you want to save, which versions to keep and for how long
- Identify milestone versions and a raw data version, which can never be altered or deleted. When working with others on data, maintain a master file
- Record the changes that are made in a new version by using a version log  
Version control can also be maintained by version-control facilities in the software you are using or in special versioning software

Versioning log				
ID-number	Who	When	What	
1	Maaïke Messelink	15-11-2014	I made the variable <i>CatReIll</i>	
2	Maaïke Messelink	16-11-2014	I changed the values of the	
3	Harrie Knippenberg	17-11-2014	I added new researchperson	

## Version control

- Git
- Subversion
- Electronic Lab Notebooks

# Versioning data

**PUBLIC SERVICE ANNOUNCEMENT:**

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:

**2013-02-27**

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13  
 20130227 2013.02.27 27.02.13 27-02-13  
 27.2.13 2013. II. 27. 27/2-13 2013.158904109  
 MMXIII-II-XXVII MMXIII <sup>LVII</sup>/<sub>CCCLXV</sub> 1330300800  
 ((3+3)×(111+1)-1)×3/3-1/3<sup>3</sup> ~~2013~~ <sup>2013</sup>   
 10/11011/1101 02/27/20/13  $\begin{matrix} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ 5 & 6 & 7 & 8 \end{matrix}$

left pad other numbers with zeros

```
01_marshall-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
```

if you don't left pad, you get this:

```
10_final-figs-for-publication.R
1_data-cleaning.R
2_fit-model.R
```

which is just sad

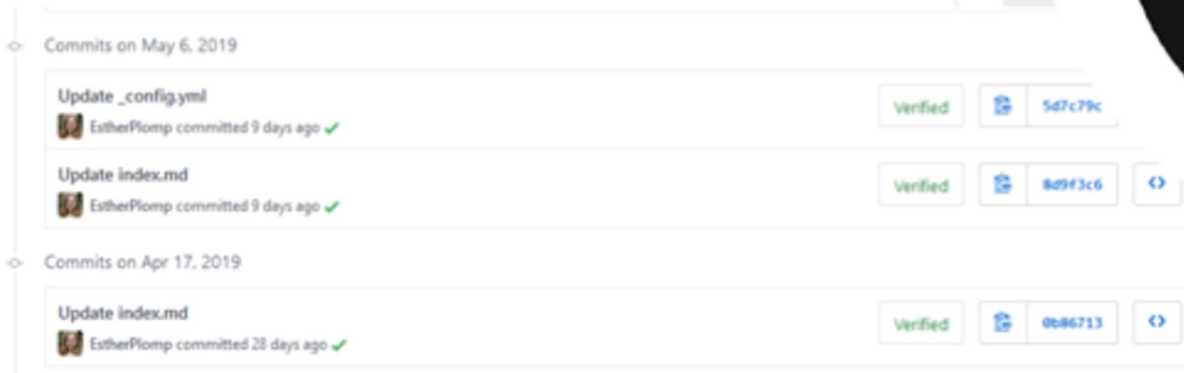
<https://speakerdeck.com/jennybc/how-to-name-files>



# Versioning data

## 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 ↔



## Summary slide: data documentation

Producing high-quality documentation in the course of your research ensures that your data can be:

- Properly interpreted as relevant context is available
- Verifiable and reproducible
- Reusable (by you or by others)

### 3 types of data documentation

- Files that explain the **content** of the dataset, at the data level (codebook)
- Files that explain the **context** of the dataset, and how the research was done (methodology section)
- Files that explain the **structure** of the dataset (readme.txt file with the structure of the dataset)

### Embedded documentation

- Code, field and label descriptions
- Descriptive headers or summaries

### Additional documentation

- Codebook (depending on the program)
- Readme file
- Methodology file
- Questionnaires or interview guides
- Working papers or lab notebooks





# Data documentation



## Lab Notebooks

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

**decrease research  
efficiency/  
reproducibility**






# Data documentation

## 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**

# Data life cycle: publishing, preserving & reusing data

PUBLISHING  
AND  
SHARING  
DATA

## Publishing and sharing data

Establish copyright; create user documentation; create discovery metadata; select appropriate access to data; publish/share data; promote data

PRESERVING  
DATA

## Preserving data

Migrate data to best format/media; store and backup data; create preservation documentation; preserve and curate data

RE-USING  
DATA

## Re-using data

Conduct secondary analyses; undertake follow-up research; conduct research reviews; scrutinize fundings; use data for teaching and learning

## Topics

- Data archiving
- FAIR software
- Data rights



## Summary slide: data archiving

Generally, RDM policies state something like this

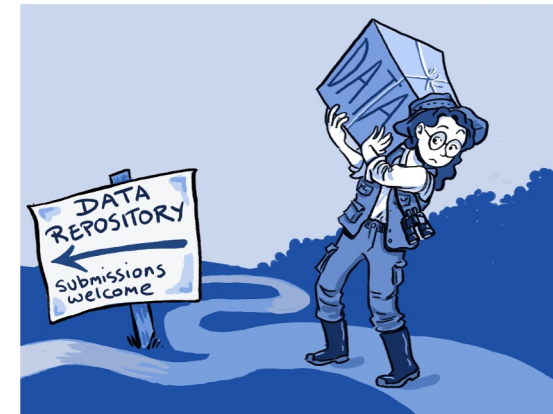
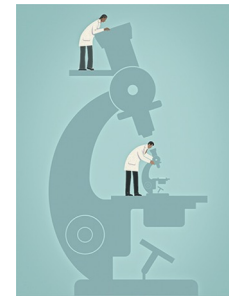
- Data are stored at the time of publication of the research (including dissertations) at the latest, together with at least all the information necessary for potential reuse of data (metadata)
- The retention period for research data is a minimum of ten years

Two perspectives

- Archiving data for scientific integrity
- Archiving data for reuse
- Rule of thumb: *as open as possible, as closed as necessary*

For consideration

- Are there ethical and legal reasons not to share my data?
- Must all data be shared?
- Where is my data safe?
- Is my data in an easy to use format?
- Will my data be accessible in the long term?
- Do I have sufficient documentation and metadata?



# Archiving data

## Archiving data for reuse

- Promote innovation and potential new data uses
- Build on each others work, which is (in most cases) funded by public money
- No duplication of data creation
- Prevent fraud and improve research integrity
- Increase visibility of research and therefore citations
- Make possible new collaborations and (possibly) publications
- Encourage scientific debate
- Meet requirements of funders, journals and universities
- Preparing data for sharing makes it also suitable for long term preservation



# Archiving data

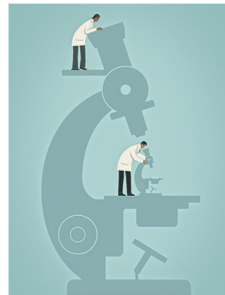
## What data should be archived?

### From the perspective of reuse

- Final (definitive) versions of data used for analysis, possibly also raw and processed data
- Documentation/codebooks necessary for understanding the data
- Read me.txt for understanding the structure and content of the deposit

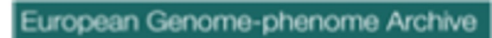
### From the perspective of scientific integrity

- Approval ethical committee
- Informed consent procedure
- Raw, processed and analysed data
- Documentation/codebooks
- Read me text
- Data management plan
- Audit trails and query trails



# Summary slide: data archiving

## Repositories



## Summary slide: data archiving

If you intend to reuse existing data instead of collecting it yourself, there are good sources for potentially relevant existing data. The following (directories of) archives may be relevant sources for finding data.

- [Open Access Directory: Data repositories](#)
- [Re3data.org](#)
- [Recommended Data Repositories from Nature](#)
- [Fairsharing](#)



**Find**

**Recommendations**

Standards and/or databases recommended by journal or funder data policies.

**1453 Standards**

Terminology Artifact Model/Format	792
Reporting Guideline	438
Identifier Schema	175
FAIR metrics	97
	18
	30

[View all](#)

**Discover**

**Collections**

Standards and/or databases grouped by domain, species or organization.

**1583 Databases**

Natural Sciences	1469
Engineering Science	308
Humanities	97
Social Sciences	155

[View all](#)

**Learn**

**Educational**

About standards, their use in databases and policies, and how we can help you.

**134 Policies**

Funder Journal Society	23
	86
	12

[View all](#)





## Archiving data

---

# Publishing and Archiving

Essential difference:

**Publishing:** Focus on visibility and accessibility of data and information.

**Archiving:** Focus on long term preservation and retrievable data and information.



Archiving data

# Publishing formats

## Data Publishing

- As supplementary data to a publication (no DOI)
- On a project website (no enough metadata)
- Via a standardised data archive (certified and curation workflow) (ideal)
- Via a domain specific data repository (self publishing, no curation workflow) (possible)
- Data paper
- Combinations



# Publishing platforms

## Overview

### Many repositories

- Multi purpose repositories like B2SHARE, public version of Figshare and Zenodo
  - Simple publishing workflow, depositor is responsible for content, quality and metadata of the published data
- Curated repositories like DANS EASY, 4TU Data centre repository, many institutional repositories
  - Depositor sends data over, then data is checked for quality to some extent
- Community-specific repositories like EBI EGA, NCBI GEO, ...
  - Expect very specific data formats
  - Have an extended data quality checking pipeline (people try to reproduce the data)



# Archiving data

## Data Repositories

Finalised Datasets  
Snapshot

Long-term  
preservation  
~10-15 years

### Findable

DOI  
Metadata

### Accessible

Control

### Interoperable

Metadata  
Vocabulary  
Open  
formats/standards

### Reusable

Licence



# Archiving data

## Alternatives to direct and open access

### Embargo

This means that the metadata for the data will be available (allowing it to be cited in related publications), but the data itself will not be made publicly accessible until the embargo has expired



### Access levels

Research data can also be shared with a restricted access. This means that the data are accessible to potential re-users only once the researcher grants access.



## When not to share your data?

**Personal data**  
anonymized.

Your research data include personal data which can or may not be

**Confidential data**  
for example

Your research data are confidential due to arrangements made with

because of the

a third (commercial) party sponsoring your research or  
confidential nature of the data.



## Summary slide: FAIR software

# FIVE RECOMMENDATIONS FOR FAIR SOFTWARE

ENDORSE

LET'S GO! →



## FAIR software

Technically, software is a special kind of data.

- Software is the result of a creative process that provides a tool for doing something.
- Software is executable, while data is not.
- Software is often built using other software.
- The lifetime of software is generally shorter than that of data. It changes and lives.

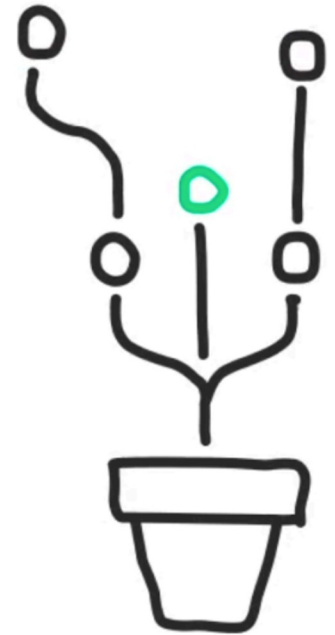




## FAIR software

# #1

**USE A PUBLICLY  
ACCESSIBLE  
REPOSITORY WITH  
VERSION CONTROL**



# FAIR software

## HELP ME CHOOSE

Git is the most feature-rich, most modern and most popular by a good margin, and we heartily recommend you use it for all you version control needs. To get the best out of Git, use it in combination with GitHub.com, Bitbucket.org, or GitLab.com.

- [How to get started with git](#)
- [Choosing a platform for your software project](#)
- [GitHub.com](#)
- [BitBucket.org](#)
- [GitLab.com](#)



# FAIR software

## #2 ADD A LICENSE



# FAIR software

## HELP ME CHOOSE

We recommend you stick to one of the more popular licenses. Because these are typically written by lawyers, the license text is precise in expressing its terms. While that carries the unfortunate side effect of being difficult to understand, the widespread use of the more popular licenses means that there is a larger number of people who understand how the letter of the law should be interpreted.

Some of our favorite licenses are the Apache-2.0 and MIT licenses. These permissive licenses have very few restrictions, allowing others to easily reuse your work. We recommend you use [choosealicense.com](https://choosealicense.com) to find out which license is best for your purposes. If you just want to check what is and what is not allowed under a given license, visit [tldrlegal.com](https://tldrlegal.com) to find out.

- [choosealicense.com](https://choosealicense.com)
- [tldrlegal.com](https://tldrlegal.com)
- [How to add a license to your GitHub repository](#)



## FAIR software

# #3

# REGISTER YOUR CODE IN A COMMUNITY REGISTRY



# FAIR software

## HELP ME CHOOSE

Community registries come in many flavors. Choosing the one that is best suited for your needs can be tricky. Here are some things to think about:

- How much traffic does the community registry get?
- Is the community registry targeting the audience you are trying to reach?
- What metadata does the community registry offer? This is sometimes described in the documentation of the registry, but you can also see for yourself by installing a tool like the [OpenLink Structured Data Sniffer](#). Alternatively, some search engines have tooling like the [Structured Data Testing Tool](#) to provide insight into how they perceive a given website.

Finally, ask a couple of colleagues which registries they would use if they were looking for software like yours.

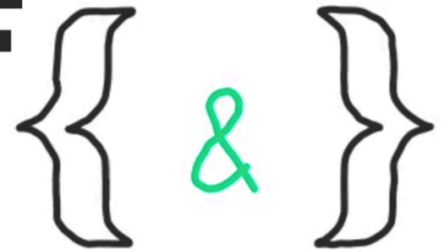
[VIEW ALL RESEARCH SOFTWARE REGISTRIES](#)



## FAIR software

# #4

# ENABLE CITATION OF THE SOFTWARE





# FAIR software

## HELP ME CHOOSE

The [CodeMeta](#) standard and the [Citation File Format](#) were specifically designed to enable citation of software and will likely meet your needs. For either one, you write a plain text file with citation metadata, which you then distribute with your software.

Initialize your CITATION.cff files [here](#).

Regarding archiving copies of your software, look for services that store their own copy of a snapshot of your software, such that whatever persistent identifier you get ([DOI](#), [URN](#), [ARK](#), etc) points to a specific version of the software, and will continue to resolve to exactly that version for the foreseeable future. Ideally, storing snapshots of your code should be as easy as possible: either at the push of a button, or automatically, for example [each time you make a new release of your software](#).

Some archiving services that meet these requirements are:

- [Zenodo](#)
- [FigShare](#)
- [Software Heritage Archive](#)



## FAIR software

# #5 USE A SOFTWARE QUALITY CHECKLIST



# FAIR software

## HELP ME CHOOSE

There are many checklists available. We find that the most useful checklist are those that

1. Allow for a granular evaluation of a software package, as opposed to just pass or fail
2. Explain the rationale behind each item in the checklist
3. Explain how to get started with implementing each item in the checklist

We recommend that you include the checklist as part of the README, for example as a badge or by including the checklist as a Markdown table. The point is decidedly not to show perfect compliance, but rather to be transparent about the state of the code while providing the necessary guidance on which aspects could be improved.

One of our favorite checklists that meets these criteria is the Badge Program developed by the Core Infrastructures Initiative, but there are many other checklists to choose from. Here is a list of some candidates:

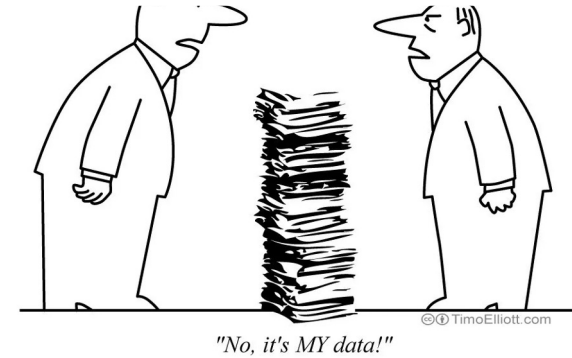
- [Core Infrastructures Initiative](#) (online, interactive)
- Deutsches Zentrum für Luft- und Raumfahrt [Class 1](#), [Class 2](#), [Class 3](#) (Markdown)
- Software Sustainability Institute's software evaluation checklist ([Google form](#))
- CLARIAH checklist ([PDF page 38-42](#))
- EURISE ([Markdown](#))



# Summary slide: data rights

## Data ownership

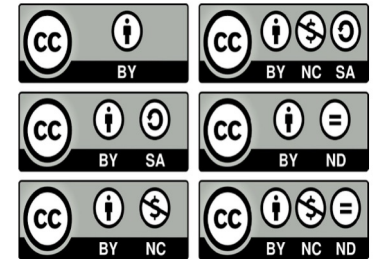
**Quite often** ... a funder, a university that pays your research, and not: you! Check your institution's policy, the funder's agreements, and the agreements made in the NUTRIOME project



**Creative Commons** licenses are also applicable to data

There are **existing licenses** that specifically apply to data. These are the so-called [Open Data Commons](#), which can be divided into three licenses:

- [Public Domain Dedication and License](#) (PDDL)
- [Attribution License](#) (ODC-By)
- [Open Database License](#) (ODC-ODbL)



Make sure that you do not hand over any author rights belonging to your data via for instance a CC0 license. In that case, your work is dedicated to the public domain by waiving all the (copy)rights. Anyone can copy, modify and distribute your data, even for commercial purposes, all without asking permission or reference to your dataset.



## Summary slide: data rights

### Data use agreements

**Data use agreements** – also known as data transfer or exchange agreements – are contracts used for the transfer of data which are non-public or otherwise subject to restrictions



Agreement between the data owner and a recipient. Composing a data use agreement may be particularly relevant in research that involves privacy-sensitive data.

[Dutch examples](#) approved by the University Medical Centers

Consider at least the following aspects:

- Legal aspects – including the General Data Protection Legislation (*Algemene Verordening Gegevensbescherming, AVG*) and the Medical Research involving Human Subjects Act (*Wet Medisch-Wetenschappelijk Onderzoek met Mensen, WMO*)
- Ownership of the data
- Privacy/anonymisation of human data
- Use of data by third parties
- Embargo period
- Citations and/or co-authorship
- Specific scientific purpose



# Data rights

## Copyrights, the owners and contracts

### *What is covered by copyright*

- Books, articles, drawings, pictures, software, databases
- NOT covered: ideas, concepts, information as such (including individual data)

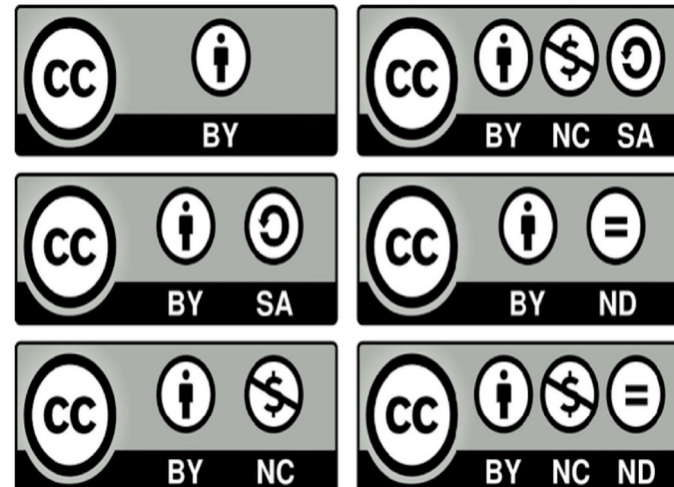
### *Basic (worldwide!) principle of copyright law*

The actual author/authors of an original work is/are the exclusive owner(s) of the right to

- Reproduce (i.e., make (digital) copies of, etc.)
- Distribute (e.g., hard copies, make available online)
- Translate

### *However, copyright can be transferred!*

The author is not entitled any more to reproduce, distribute, translate, etc ...



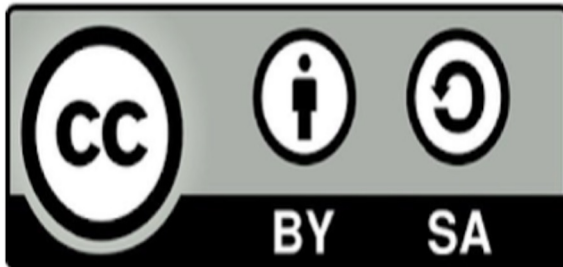
<https://creativecommons.org/>



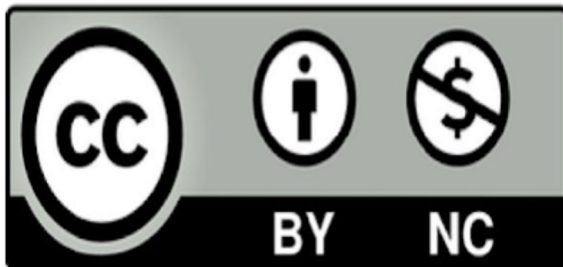
## Data rights



ATTRIBUTION: All CC licenses require that others who use your work in any way **must give you credit the way you request**, but **not in a way that suggests you endorse them or their use**. If they want to use your work without giving you credit or for endorsement purposes, they must get your permission first.



+SHARE ALIKE: You let others copy, distribute, display, perform, and modify your work, as long as they **distribute any modified work on the same terms**. If they want to distribute modified works under other terms, they must get your permission first.



+NON-COMMERCIAL: You let others copy, distribute, display, perform, and (unless you have chosen NoDerivatives) modify and use your work **for any purpose other than commercially** unless they get your permission first.

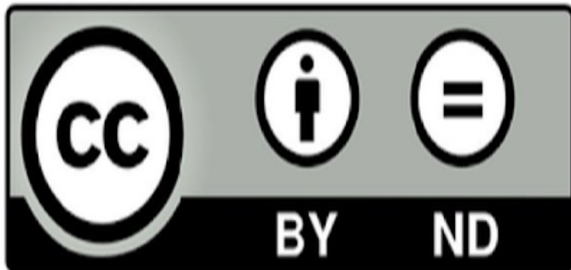




## Data rights



**Combines SA and NC** (users should distribute any modified work on the same terms, and can use your work for any purpose other than commercially)



+NON-DERIVITAVES: You let others copy, distribute, display and perform **only original copies of your work**. If they want to modify your work, they must get your permission first.



**Combines NC and ND** (users are only allowed to use original copies of your work and for any purpose other than commercially).

