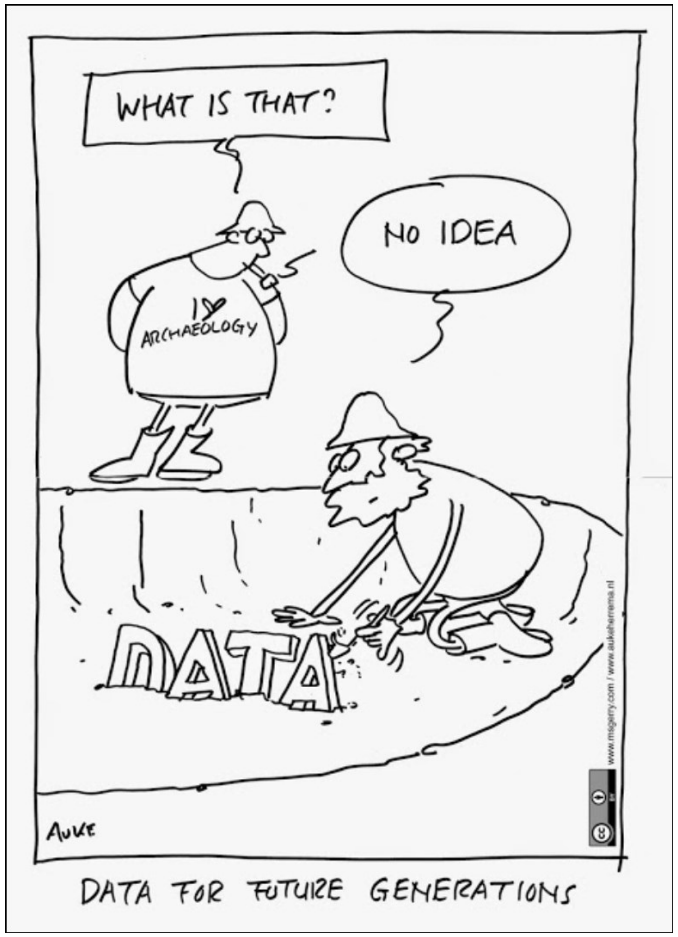
Quick Start to Research Data Management

Team RDM Support

Outline



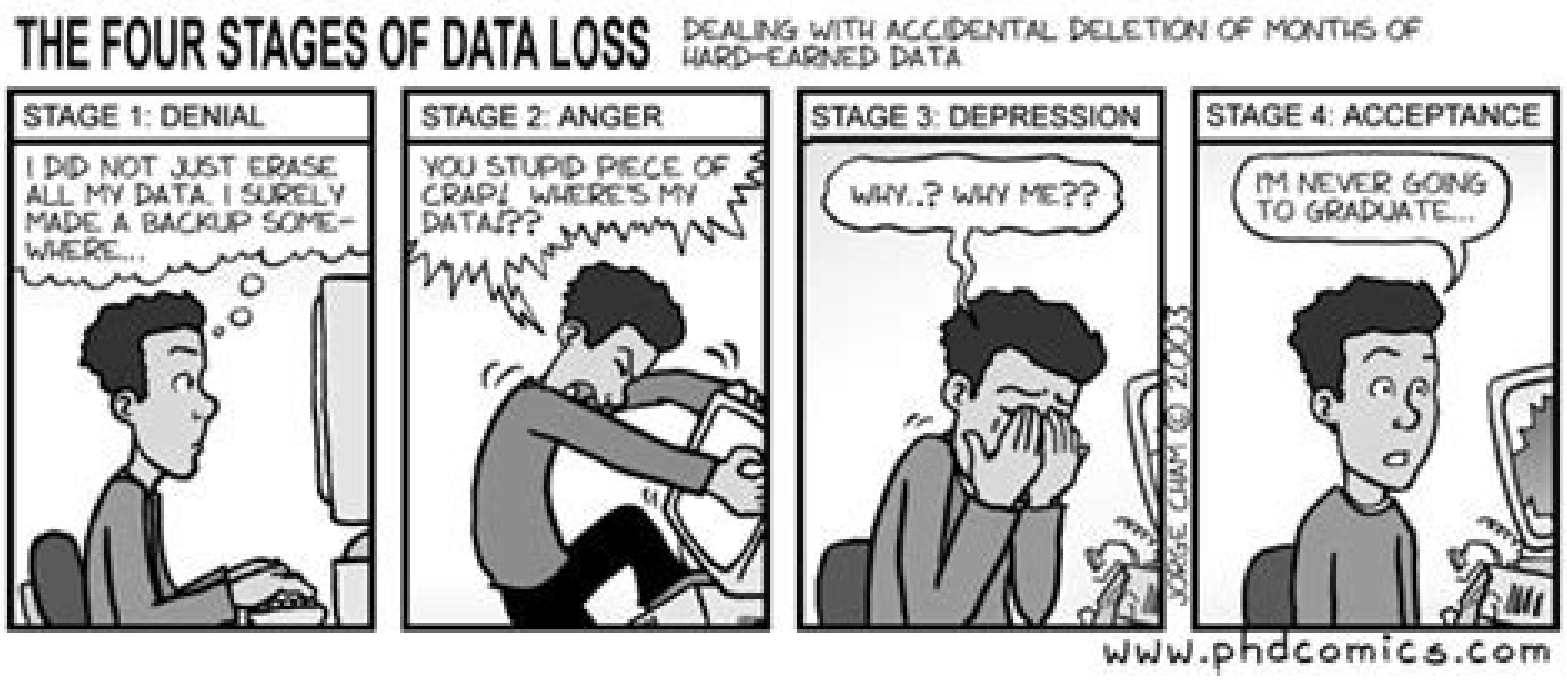
- Why is Research Data Management (RDM) important?
- FAIR vs open data
- Data Management Plan (DMP)
- Data collection
- Data documentation & metadata
- Data storage
- Data privacy & security
- Data selection & archiving
- Data sharing & reusability

Utrecht University

Why is RDM important?

Why is RDM important?

- Data is valuable
- Data is fragile and easily lost
- Improves research efficiency and quality
- Facilitates collaborative research



Why is RDM important?

- Reproducibility
  Data can be reused by yourself or by other research groups
- Research integrity
  Verification & transparency
- Funder mandates and policies
  NWO, ERC, Horizon, ZonMW require DMP for every project
- Legislation
  Personal data must abide by GDPR regulations

Read more about research integrity at the UU on our website Research integrity - Research - Utrecht University (uu.nl)

Utrecht University

FAIR vs open data

FAIR principles

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

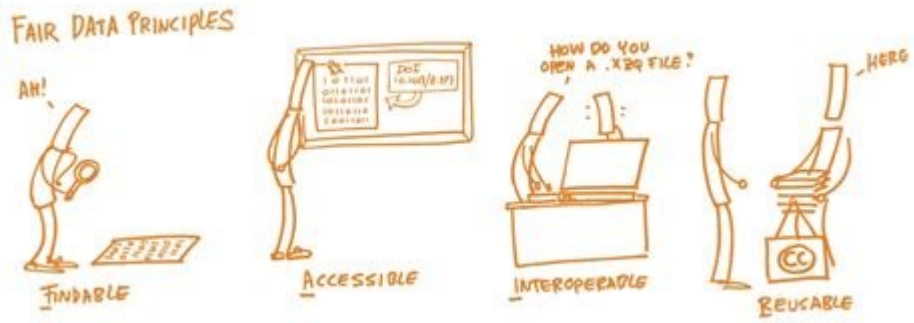
ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE



https://www.openaire.eu/images/Guides/FAIRdataprinciples_foster.png

What is FAIR DATA? Source: Ligue des Bibliothèques Européennes de Recherche, CC-BY

FAIR vs. open

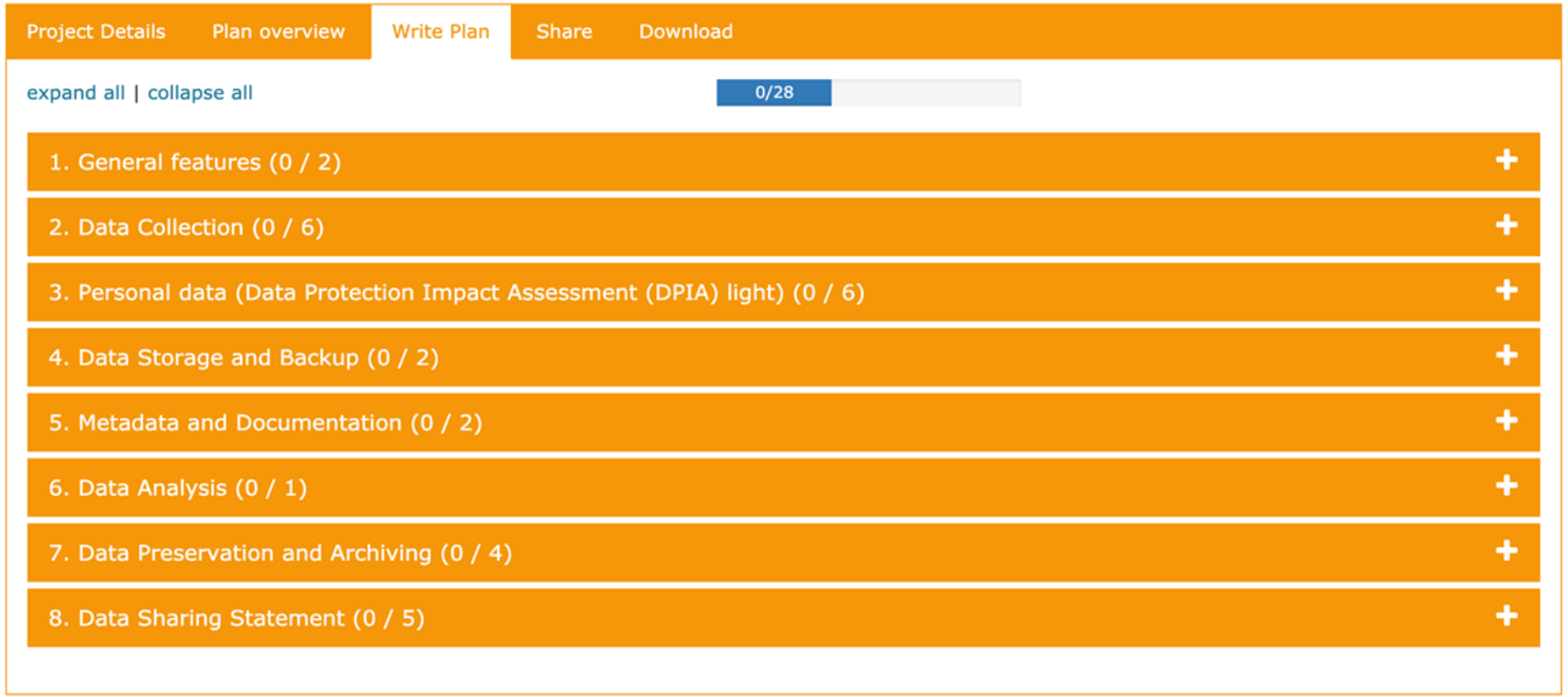
“Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share-alike.”
– Open Data Handbook

Make your data… As open as possible as closed as necessary“
(European Commission)

Utrecht University

Data Management Plan (DMP)

Data Management Plan (DMP)



Use your institutional credentials to log into DMPonline: https://dmponline.dcc.ac.uk/

Utrecht University

Data collection

Data
Preferred formats

What is data?

Models

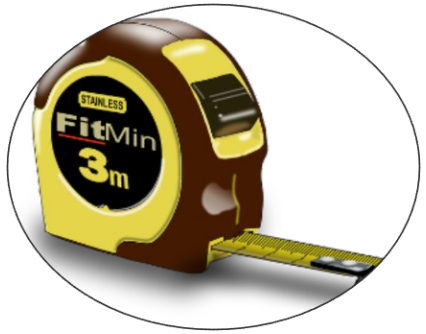
MRI

Geographical data

Survey

Literature

Musical Score



Measurements

Interviews

Library reference

Data collection

- What type of data?
- Which formats?
- How many files?
- What size of files?
- Origin of data?

Data collection table

Type	Description	Origin /collection	Formats	Software	Total file size	Number of files / samples
Lab and stable journals	Dates, protocols, lab worker, etc.	Lab worker / researcher	.csv and .txt	eLabjournal	100-500 MB	2 lab journals (consist of multiple files)
Biological data	Blood samples	Veterinarian	1mL/animal	NA	NA	250 animals
Lab results	Gene expression and antibody titers	Microarray data and ELISA data	.csv, .Rdata, .chp, .txt	Affymetrix, locally developed tool	200 GB	20 data output files
Behavioural data	Animal behavior visually scored	Researcher and research assistants	.csv	Noldus observer and ethovision	Kb's	2 output datafiles
Bodyweight	Biweekly bodyweight	Stable workers	.csv	NA	Kb's	1 output datafile
Statistical analyses	Scripts/codes and output tables and figures	Researcher	.R, .SAS, .csv, .tiff	R, Rstudio, SAS, Excel	1-50 MB	5 scripts, 5 table files, 5 figure files

Data formats

Preferred formats

Text

Quantitative data

Images

Audio

Video

Preferred formats offer the best long term guarantees in terms of usability, accessibility and sustainability.

Preferred Formats Characteristics:

- Non-proprietary / Open Source
- Unencrypted
- Uncompressed
- Interoperable
- And/Or: Commonly used

https://dans.knaw.nl/en/file-formats/

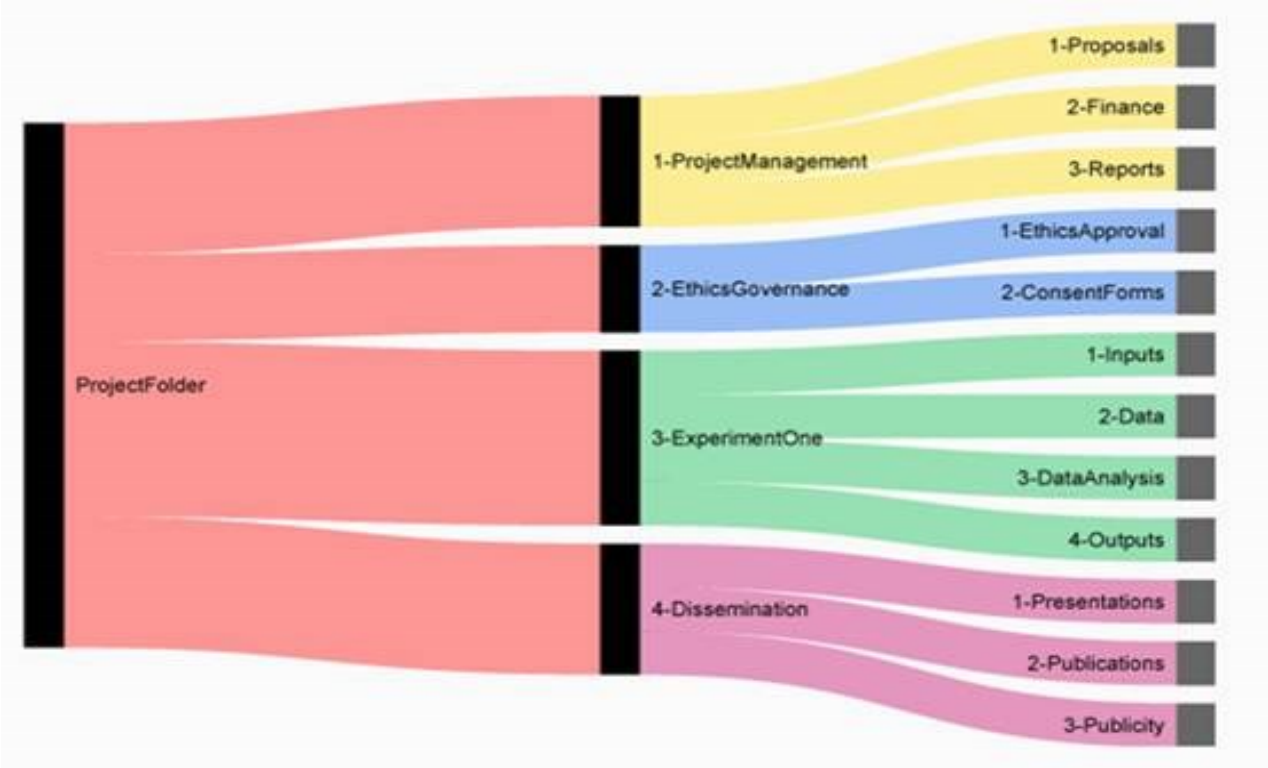
Utrecht University

Data documentation

Folder structure and file naming
Version control
Metadata and documentation

Does your desktop look like this?

It may be time for some structure!



Example folder structure
From: 'Setting up an Organised Folder Structure for Research Projects' Posted June 4, 2014 Blog by Nikola Vukovic

Folder structure tips

- Create folder hierarchies based on:
  - File formats
  - Raw vs processed data
  - Date
  - Experiments
- Determine in advance what your folder structure should be
- Make it logical
  - From time to time, ‘check the logic’
- Describe your structure and files in a README file

Folder “tree”

You can easily create a folder “tree” to use within your read_me file or DMP by using the command prompt command “tree”.

To do so:

1. Navigate to your desired folder using the “cd” command.
2. Use the command “tree” when at the desired folder.
3. Copy-paste the folder tree.

Research project package

Part of documentation is also determining the contents and organization of the Research Project Package:

- **Data**
  - Raw data
  - Processed data
- **Documentation and Metadata**
  - Variable codebook
  - Metadata standard (At the data and descriptive level; if any)
  - Protocols and Standard operating procedures (SOP)
- **Scripts & Software**
  - Analysis
  - Processing
- **Legal Documents**
  - Data protection impact assessment (DPIA)
  - Licenses
  - Informed Consent form (template only)
  - Data transfer agreement (DTA)
- **Administration**
  - Ethical review
  - Grant & Consortium: applications and agreements
  - Data management plan (DMP)

Folder structure tools

File Renaming Tools

- Bulk Rename Utility (for Windows)
- Renamer (for MacOS)
- PSRenamer (for MacOS, Windows, Unix, Linux)
- WildRename (for Windows)

File naming do's

When constructing a file name, consider using one or more of the following:

- Project or experiment name or acronym
- Location/spatial coordinates
- Researcher name/initials
- Date or date range of experiment
- Type of data
- Conditions
- Version number of file

This will differ depending on your project/experiment. Find a method that fits best for your data. There is no "perfect method"!

File naming don’ts

Avoid the following when naming files

- Special characters (&%$#/)
- Periods
- > 25 characters
- DAY-MONTH-YEAR (better YEAR-MONTH-DAY)
- Duplicate files in different folders
- Whitespace (better ‘-’ OR ‘_’)

File naming quiz question

Which of the following filenames has all do's and no don'ts?

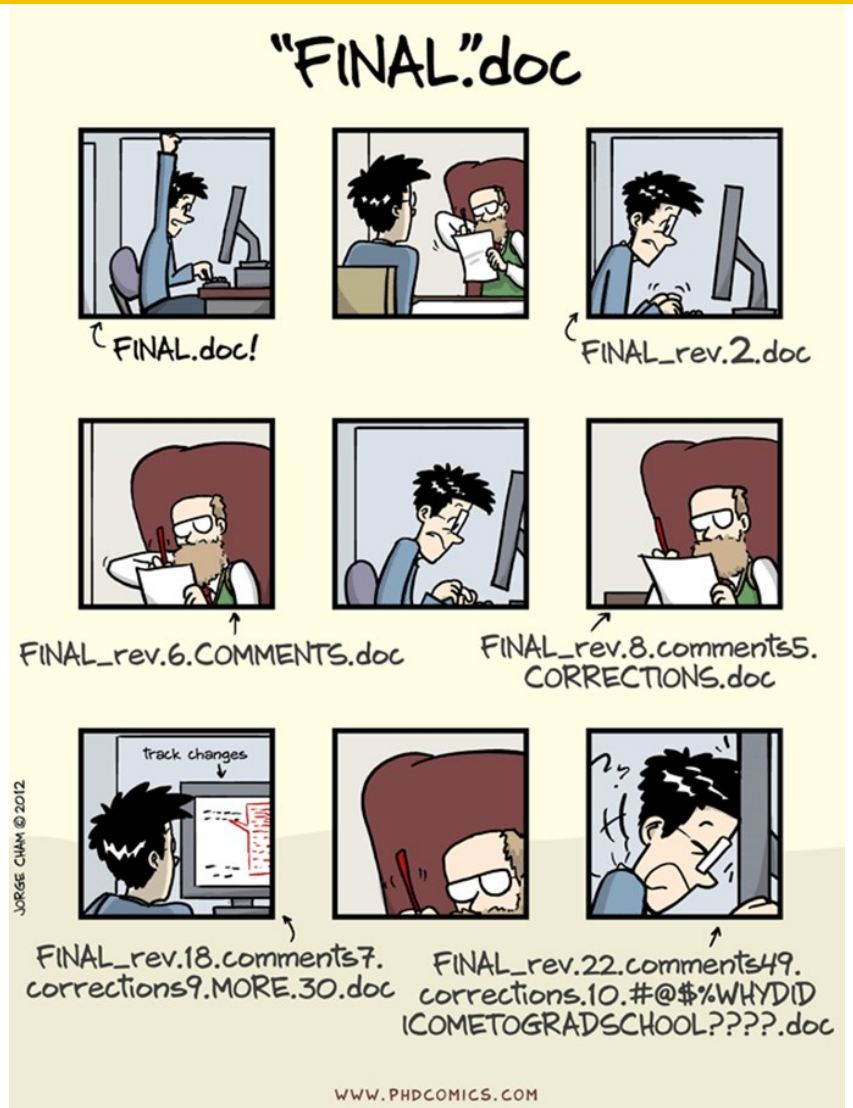
A. MA_NTC023_20141031.xls
B. MA@NTC#23~20141031.xls
C. MicroArrayData_NetherlandsToxicogenomicsCentreProject023_20141031.xls
D. microarrayntc02320141031.xls
E. MA_NTC023_31102014.xls
F. MA/NTC/Project23/OCT31st/data.xls

Folder structure and file naming recap

Remember your folder structure and file naming have to make sense:

- ✓ to you
- ✓ to you - after the holidays
- ✓ to you - after graduation
- ✓ colleagues within a project
- ✓ project leaders / supervisors when contract ends
- ✓ etc.

Version control



Version control is the process of managing changes to files.

Manual version control can be as simple as adding a version number or a date (using YYYYMMDD format) to the end of the file name.

Examples:
ASIST_abstract_v3.docx
ASIST_abstract_20180522.docx

Source https://simmons.libguides.com/c.php?g=814790&p=5983200

Metadata and documentation

Make data usable for others

Metadata Structured data providing information about one or more aspects of the data

Documentation

Creator: Peter Parker
Place: Stellendam
Date: 10.10.2018
Diafragma: 3.2
Lighting: 4.3
Source: Wikipedia
Description:
Pr02_urban_2015.txt

Documentation

All contextual information pertaining to your data, data files

- Methodology (i.e methods section of an article)
- Data management plan
- Codebooks, controlled vocabularies (data level metadata)
- READ_ME.txt file
- Lab book
- Etc…

It is a crucial aspect of interoperability which allows others to use your work!

Metadata – Data about data

Metadata exists at different levels:

- **“Data level” metadata** describes variables directly related to your dataset
  - Rewards (Number of sugar pellets earned by the rat in the task)
  - Drug (Name of administered analgesic (only two options: Paracetamol or Ibuprofen))

  i.e. (codebook, controlled vocabularies, microscope settings, interview context …)

- **“Descriptive level” metadata** describes higher order aspects of your dataset
  - Author
  - Affiliated institution
  - Language

Metadata standards

Data level

Metadata standards exist for certain disciplines and fields

- They are structured vocabularies that facilitate the exchange and comparison of data in identical subjects;
- This prevents issues wherein identical variables are called differently and categorized separately when they in fact point to the same type of observation;
- Rewards / sugar collected / pellets earned = all equal rewards but may cause confusion when no metadata standard has been defined

Metadata Standards

Descriptive level

- Descriptive Metadata Standards are quite common.
- These are often defined by the repository where you will be preserving or sharing your data.
- Common ones are: **Datacite 4.0 & Dublin Core**

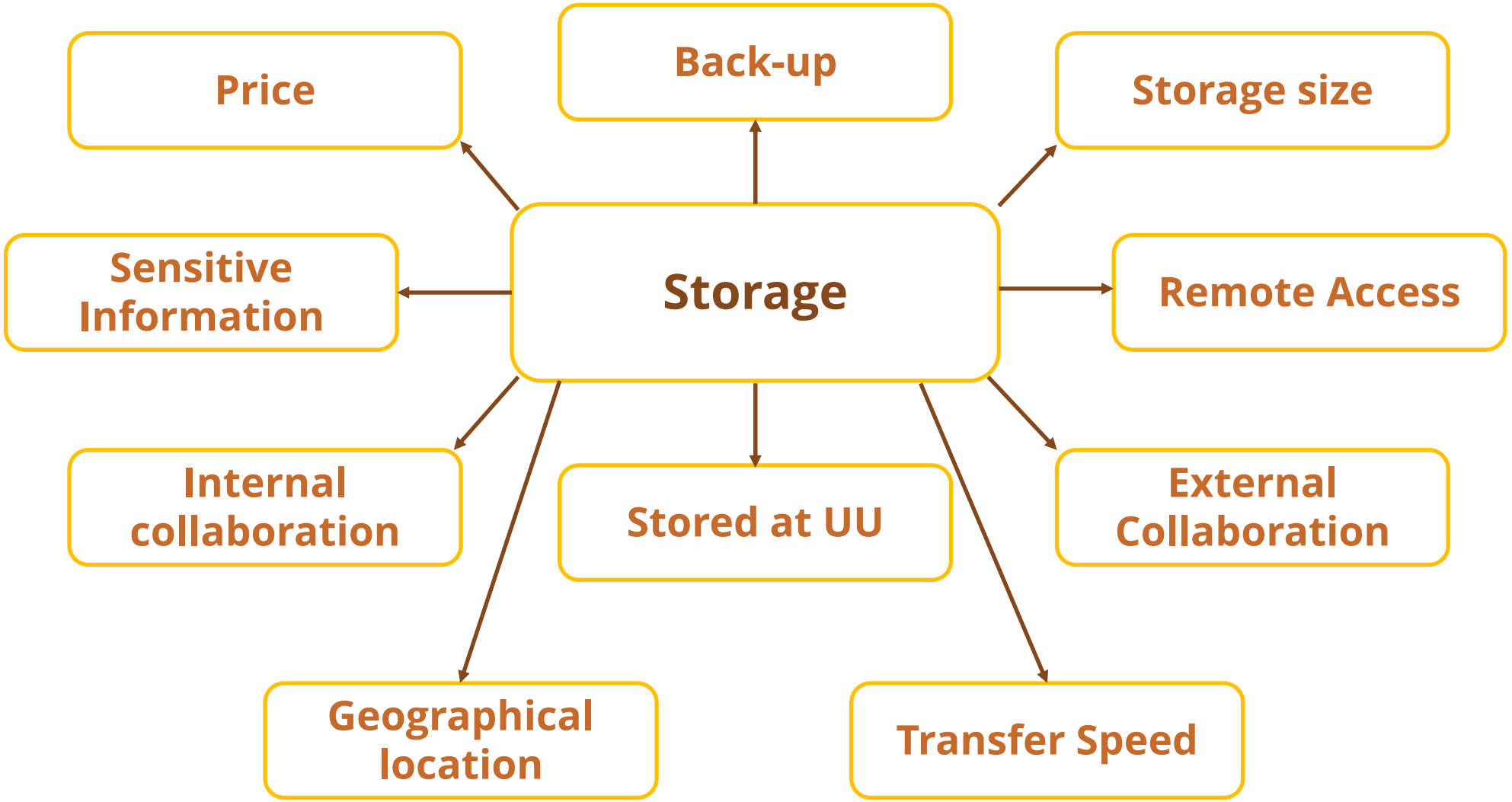
Dataverse and Yoda both use Datacite 4.0 as their descriptive metadata standard.

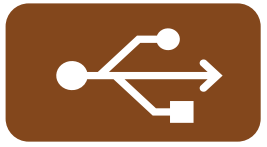
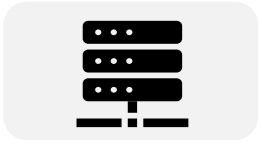
Utrecht University

Data storage

Choosing storage options
UU storage solutions

What do you consider while choosing a storage solution?



Storage Option					YODA	OneDrive	SURF DRIVE
Storage size	Varied	Varied	Varied	Varied	Varied	1TB	250GB
Price	NA	NA	Faculty	Faculty	TB €4/m	UU	UU
Back-up	✗	✗	✓	✓	✓	✓	✓
Controlled by UU	✗	✗	✓	✓	✓	✓	✓
Internal collaboration	✗	✗	✗	✓	✓	✓	✓
External Collaboration	✗	✗	✗	✗	✓	✓	✓
Sensitive Information	✗	✗	✓	✓	✓	✓	✓
Remote Access	✗	✗	✓	✓	✓	✓	✓

Storage solutions

Please use this tool to select an appropriate solution:

Tools | Data Storage Finder (uu.nl)

Read more about information security at the UU on our intranet: Information Security - Safe tooling and software - Information security at the UU - Intranet

Utrecht University

Data privacy & security

Personal data
Data security
Privacy and security measures

Personal data: the 7 principles of the GDPR

Definitions

Data Subject

- The individual whose data is being collected
- Natural Person

Personal Data

- Any information which can identify a natural person whether **directly or indirectly**

Sensitive Personal Data

- Personal information revealing a data subjects' ethnic origin, political opinions, religious beliefs,
  sexual orientation and genetic data or biometric data uniquely identifying a data subject

Definitions

Processing

- **ANY** operation or set of operations that is performed on personal data

Controller

- The natural/legal person or other body which, alone or jointly with others, **determines the purposes and means of the processing** of personal data

Processor

- The natural/legal person or other body which processes personal data **on behalf of the controller**

Controller quiz question

Who is the controller of the personal data you work with ?

Is it:

- The master student who collected it ?
- The Phd student overseeing the project?
- The Primary investigator?
- The Dean of the Faculty
- The University?

Controller quiz question answer

In most cases the **University you work for is the Controller**. You as an employee of the University oversee the compliance of your study, but you do so as the University, not as an individual.

The GDPR carries with it hefty fines in case of non-compliance. If something bad occurred in the handling of the personal data of your participants and a lawsuit would arise. It would be the University who is held responsible and sued and not you as a researcher.

In other cases, the University may not be the sole controller. If you obtained the data from a tertiary source, this third party may also be a Controller.

Thus, there can be more than one controller !

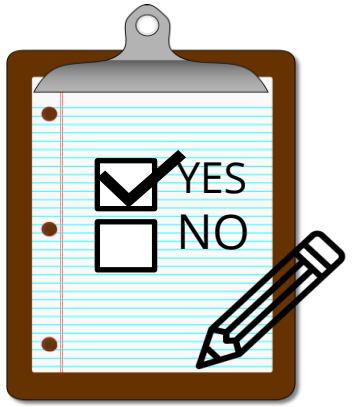
A controller is simply someone who determines the means of processing of personal data. Any institution or body who handles, processes and determines how and what will be done with the data, automatically becomes a controller (even if they are unaware of it!).

Lawfulness of Processing (Art. 6)

Personal data may only be processed if at least one of the following applies:

Most commonly used:

1. **Informed Consent**



Less commonly used:

2. **Legitimate interest of the controller**
3. **Public Interest**

Hardly ever used in research:

4. Legal Obligation
5. Contractual
6. Vital interest of the data subject

Informed Consent

✓Freely given

Must be a real choice and not influenced by external factors

✓**Specific**

Bound to several specified purposes which are sufficiently explained

✓**Informed**

What kind of data; How it will be used; With what purpose; Right to withdraw

✓Unambiguous

A clear affirmative statement

Informed Consent

Used as a means to meet the
Legal and **Ethical** obligations
a researcher holds towards
their participants

Informed Consent: information sheet

Data subjects must be (at the very least) informed about:

- The controller's identity and contact details (primary contact)
- DPO's contact details (privacy@uu.nl) (secondary contact)
- Purpose of research
- Legal basis for collecting their personal data
- Personal data being collected
- Right to withdraw consent and how to do so.

Other requirements may be in place for

- Third country transfers
- Multiple controllers
- Automated Decision-making processes

Data security: three levels

Security of data files

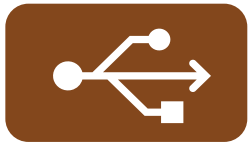
- Encryption
- Pseudonymization
- Minimization
- Abstraction
- Anonymization
- Access control
- Secure transport
- Complete deletion

Security of computer system

- Firewall
- Antivirus software
- Installing updates
- Using secured WIFI
- Password-protected
- Device encryption

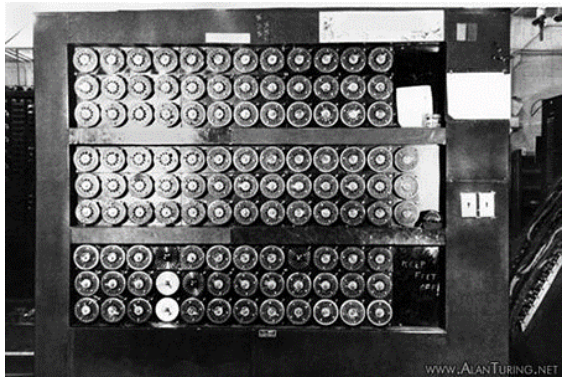
Physical data security

- Key & Lock
- Don’t leave unattended
- Safe transport

Storage Option					YODA	DataverseNL	OneDrive	SURF DRIVE
Storage size	Varied	Varied	Varied	Varied	Varied	Varied	1TB	250GB
Price	NA	NA	Faculty	Faculty	TB €4/m	UU	UU	UU
Back-up	✗	✗	✓	✓	✓	✓	✓	✓
Controlled by UU	✗	✗	✓	✓	✓	✓	✓	✓
Internal collaboration	✗	✗	✗	✓	✓	✓	✓	✓
External Collaboration	✗	✗	✗	✗	✓	✓	✓	✓
Sensitive Information	✗	✗	✓	✓	✓	✓	✓	✓
Remote Access	✗	✗	✓	✓	✓	✓	✓	✓

Privacy and security measures

Encryption



Pseudonymization

Anonymization

Encryption

Encryption makes data unreadable/inaccessible to those without a password

Different levels of encryption

- Data encryption – the **data file** itself is encrypted
- Drive encryption – the **hard drive,** where the data file is located, is encrypted

Encryption tools

VeraCrypt

FileVault

BitLocker

Recommended

boxcryptor

https://intranet.uu.nl/en/information-security-how-do-i-handle-usb-sticks-and-portable-hard-drives

Pseudonymization

Replace personal identifiers within a database with an artificial pseudonym or code.

Separate the pseudonym-personal identifier key and store it separately.

Bruce Wayne	→	Batman
6’2		6’2
95kg		95kg
1007 Manor street, Gotham	→	X09785

Key Stored Separately
Batman = Bruce Wayne
X09785 = 1007 Manor street

Minimization

Removing personal identifiers within a database

Bruce Wayne	→	Batman
6’2		N/A
95kg		95kg
1007 Manor street, Gotham		X0978512

Aggregation and abstraction

Aggregating personal data within larger bins

Reducing granularity

Bruce Wayne		Batman
6’2		N/A
95kg	→	Over 90kg
1007 Manor street, Gotham		X0978512

Restricted access

- During the research
- After the research

Who should have access to your data?
How will you restrict access?
How will you enable access to those authorized?
Where will you describe who gets to access the data?

More data privacy:

If you would like to know more about data privacy in research please check out the **Data Privacy Handbook**:

https://utrechtuniversity.github.io/dataprivacyhandbook/

And our **Handling Personal Data in Research Workshop**

Utrecht University

Data selection & archiving

Research project package

- **Data**
  - Raw data
  - Processed data
- **Documentation and Metadata**
  - Variable codebook
  - Metadata standard (At the data and descriptive level; if any)
  - Protocols and Standard operating procedures (SOP)
- **Scripts & Software**
  - Analysis
  - Processing
- **Legal Documents**
  - Data protection impact assessment (DPIA)
  - Licenses
  - Informed Consent form (template only)
  - Data transfer agreement (DTA)
- **Administration**
  - Ethical review
  - Grant & Consortium: applications and agreements
  - Data management plan (DMP)

Data archiving package

- **Data**
  - Raw data
  - Processed data
- **Documentation and Metadata**
  - Variable codebook
  - Metadata standard (At the data and descriptive level; if any)
  - Protocols and Standard operating procedures (SOP)
- **Scripts & Software**
  - Analysis
  - Processing
- **Legal Documents**
  - Data protection impact assessment (DPIA)
  - Licenses
  - Informed Consent form (template only)
  - Data transfer agreement (DTA)
- **Administration**
  - Ethical review
  - Grant & Consortium: applications and agreements
  - Data management plan (DMP)

Data publication package

- **Data**
  - Raw data
  - Processed data
- **Documentation and Metadata**
  - Variable codebook
  - Metadata standard (At the data and descriptive level; if any)
  - Protocols and Standard operating procedures (SOP)
- **Scripts & Software**
  - Analysis
  - Processing
- **Legal Documents**
  - Data protection impact assessment (DPIA)
  - Licenses
  - Informed Consent form (template only)
  - Data transfer agreement (DTA)
- **Administration**
  - Ethical review
  - Grant & Consortium: applications and agreements
  - Data management plan (DMP)

Data archiving duration

UNIVERSITY POLICY FRAMEWORK FOR RESEARCH DATA (UU)

Raw data from research must be kept available for
a minimum of ten years for verification purposes

...commencing from the date that the research results are published!

Read more about this in the policy framework:
university_policy_framework_for_research_data_utrecht_university_-_january_2016.pdf (uu.nl)

Utrecht University

Data sharing & reusability

Archiving vs publication
Making data available
Licenses

Archiving & Publication

Archiving data for future reference

- Data archiving package
- Make sure you can read and access the data later on.
- Access to others for verification purposes.

*"Archived research data are to be retained for a minimum of **ten years**, commencing from the date that the research results are published." (UU)*

Publishing data for reuse

- Data publication package
- Making (information about) data usable for others.

"NWO introduced its research data management policy which aims to make research data generated as part of NWO funded projects as open and FAIR (findable, accessible, interoperable & reusable) as possible."

Making data available

To take into account:

1. Persistent identifier (PID)
2. Guarantees for sustainable availability
3. Quality control
4. Costs
5. Physical storage location

Please use this tool to select an appropriate solution:
Tools | Data Repository Finder (uu.nl)

Read more about information security at the UU on our intranet: Information Security - Safe tooling and software - Information security at the UU - Intranet

Data repository options

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

- **A domain repository**
  Use a trusted repository already established for your research domain.

- **An institutional or recommended data repository**
  If a domain repository is not available, use an institutional research data repository

- **A general purpose repository**
  If none of the above is available, use a general purpose repository like Dans EASY, 4TU.Center, Zenodo, Figshare.

recommendations by OpenAIRE (2016b)

Data Journals

ELSEVIER

Data in Brief

Editors-in-Chief: Hao-Ran Wang, Ganhui Lan

View Editorial Board

DATA SCIENCE JOURNAL

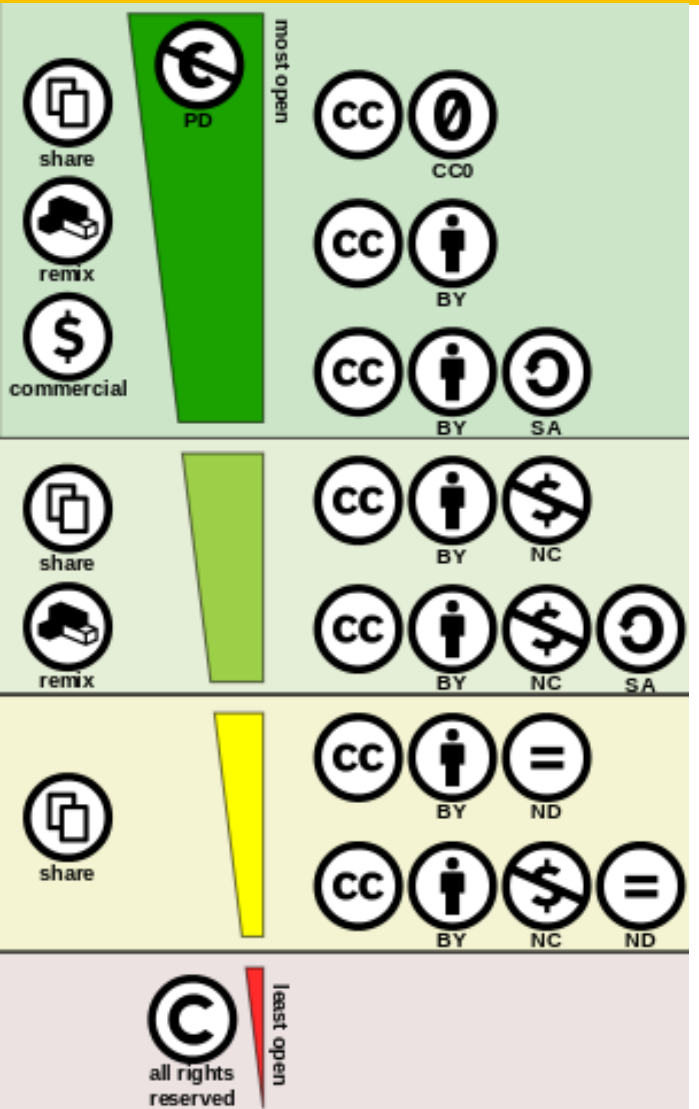
BMC Part of Springer Nature

BMC Research Notes

For high-quality datasets consider publishing a data paper in a data journal:

1. Includes an element of peer review of the dataset
2. Increase data visibility and chances of being re-used
3. Get academic accreditation

Licensing of data



A license states what a user is allowed to do with your data and creates clarity and certainty for potential users.

Creative common licences are easy to use but not the most suitable for data

CC0 No Rights Reserved

CCBY Attribution: cite the source at reuse

Source: Creative commons Nederland http://creativecommons.nl/eerste-keer/#gebruiken; http://creativecommons.org

Licensing of data

Custom License

To adequately protect your data from unwanted use it is recommended to use a custom license instead of a creative commons license.

This requires you to make your data available under restricted access and only grant access after a contract with the adequate terms and conditions has been signed by the requestor.

This very commonly used when dealing with personal data (Data Transfer Agreement).

Utrecht University

Sharing science,
shaping tomorrow

Research Data Management Support (RDM)

Got a question? Need to consult an expert?

- Data publishing (YODA, DataverseNL and other repositories)
- Data storage (YODA, One-drive, Surf-drive, research drive, U/O: drive …)
- Data management plans (DMP)
- Scripting
- High performance computing
- Machine Learning
- Privacy (GDPR, informed consent)
- GitHub
- R and Data, Python and Data
- …anything data related!

Contact us at: info.rdm@uu.nl
Website: www.uu.nl/rdm