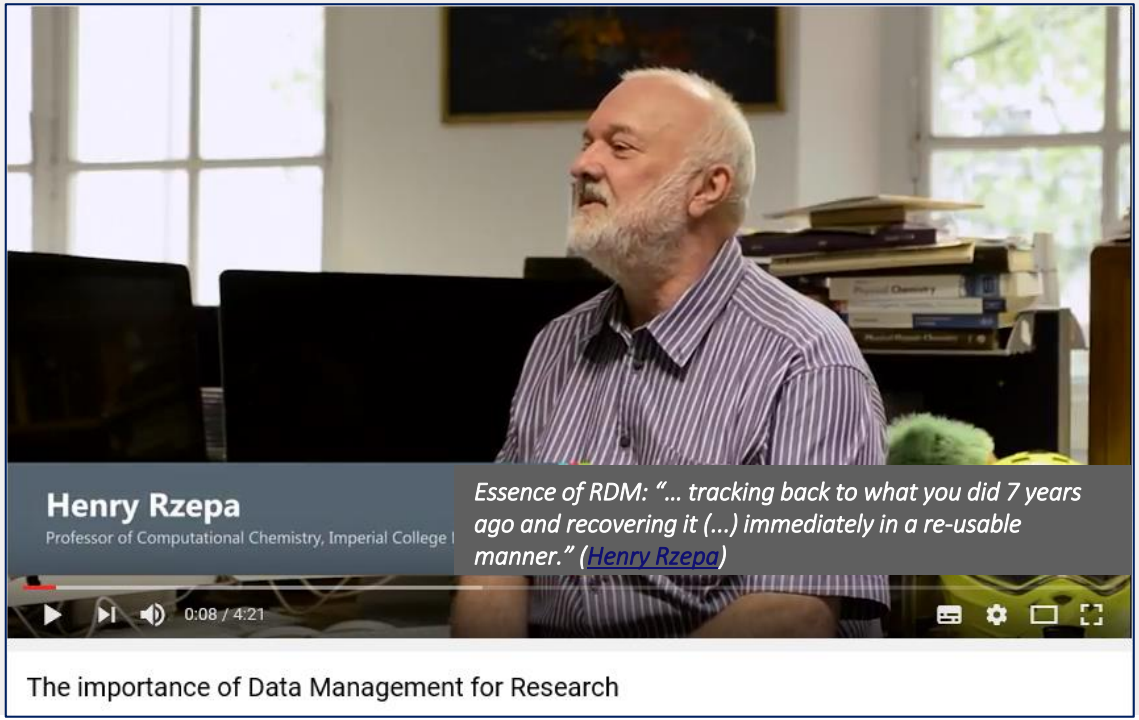
PROOF course Open science: the new default in science, 01-10-2019

Part Research data management

Leon Osinski

Information Expertise Center

1. Why research data management?

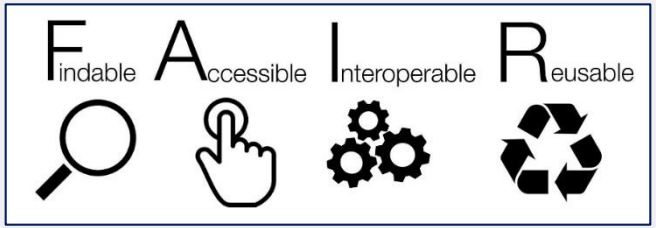


1. What is research data management?

RDM: caring for your data with the purpose to:

1. protect their mere existence: data loss, data authenticity (RDM basics)*
2. share them with others
    a. for reasons of reuse: in the same context or in a different context; during research and after research
    b. for reasons of reproducibility checks → scientific integrity; data quality; data provenance*

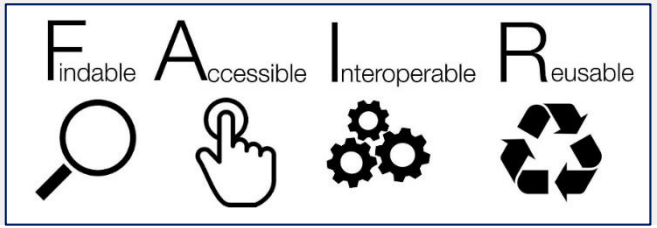
RDM prepares for data sharing → data practices that make your data findable and available to, and understandable and easy to work with for humans and machines



* Green: not mentioned by Henry Rzepa

Outline

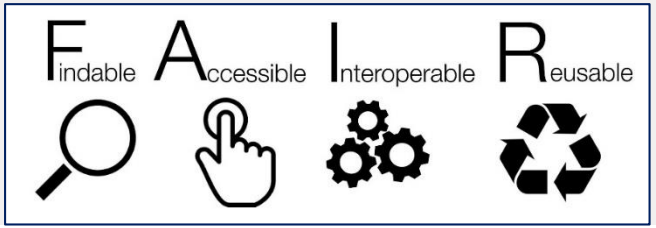
1. Research data management [RDM]: what and why
2. Caring for your data, or making your data (Re-)usable and Interoperable
    a. tidy data
    b. metadata/documentation
    c. licenses
    d. open data formats
3. Sharing your data, or making your data Findable and Accessible
    a. data protection: back up, file naming, organizing data
    b. data sharing: collaboration platforms, data archives



TU/e

RDM part 2: (re-)usable data

1. Research data management [RDM]: what and why
2. *Caring for your data, or making your data (**R**e-)usable and **I**nteroperable*
    a. *tidy data*
    b. *metadata/documentation*
    c. *licenses*
    d. *open data formats*
3. Sharing your data, or making your data **F**indable and **A**ccessible
    a. data protection: back up, file naming, organizing data
    b. data sharing: collaboration platforms, data archives



TU/e

2. Making your data (re)-usable: examples of bad data

#otherpeoplesdata

Humayun, M., e.a., Origin and age of the earliest Martian crust from meteorite NWA 7533. https://dx.doi.org/10.1038/nature12764

Nature Magazine tends to publish fabulous cutting-edge scientific research data of different types bundled all together in a PDF called "supplementary information".

In this example, they have bundled together:

- words
- image data
- scatterplot data
- a bar chart
- some awful sideways printed tables of numbers

… some say this is one of the world's "best" research journals.

2. Making your data (re)-usable: examples of bad data

#otherpeoplesdata

Findable and accessible but not usable

Humayun, M., e.a., Origin and age of the earliest Martian crust from meteorite NWA 7533. https://dx.doi.org/10.1038/nature12764

Nature Magazine tends to publish fabulous cutting-edge scientific research data of different types bundled all together in a PDF called "supplementary ...tion".

...ve bundled together:

- words
- image data
- scatterplot data
- a bar chart
- some awful sideways printed tables of numbers

... some say this is one of the world's "best" research journals.

TU/e

Data is usable when a machine can easily process it and humans can understand it

2.a. Tidy data

Tidy data allow your data to be easily processed by computers, i.e.:

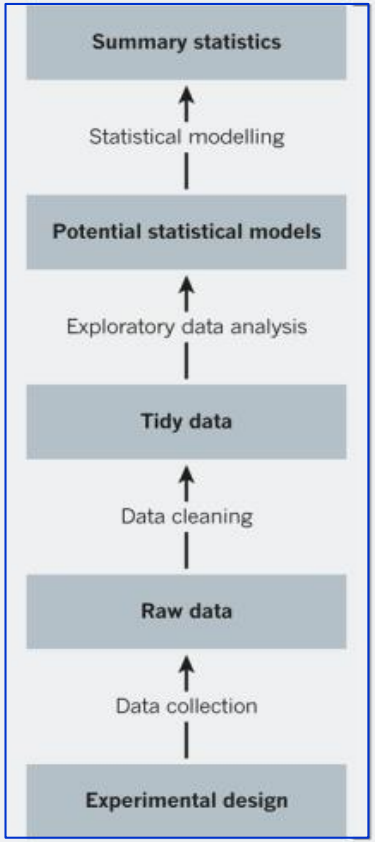
- imported by data management systems
- analyzed by analysis software
- visualized, modelled, transformed
- combined with other data (interoperability)

Tidy data

Tidy data is about structure of a table / data set.

Tidy data ≠ clean data. It's a step towards clean data

- Each variable you measure is in one column
- Column headers are variable names
- Each observation is in a different row
- Every cell contains only one piece of information



Nature, vol. 520, 30 April 2015, p. 612
http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412

Tidy data / by Hadley Wickham, http://dx.doi.org/10.18637/jss.v059.i10

Tidy data

1. Each variable you measure is in one column
2. Column headers are variable names
3. Each observation is in a different row
4. Every cell contains only one piece of information

Messy data

1. More than one variable in a single column (‘clumped data’)
2. Column headers are values, or: one variable over many columns (‘wide data’)
3. Variables are in rows and columns
4. More pieces of information in one cell (cells are highlighted or coloured; values and measurement units in one cell)

Wide data: one variable over many columns

patient_id	drug_a	drug_b
1	67	56
2	80	90
3	64	50
4	85	75

Tidy data

patient_id	drug	heart_rate
1	a	67
2	a	80
3	a	64
4	a	85
1	b	56
2	b	90
3	b	50
4	b	75

By Economic Status and Sex

Economic Status	Population Exposed to Risk			Number of Deaths			Deaths per 100 Exposed to Risk		
	Male	Female	Both	Male	Female	Both	Male	Female	Both
I (high)	180	145	325	118	4	122	65	3	37
II	179	106	285	154	13	167	87	12	59
III	510	196	706	422	106	528	83	54	73
Other	862	23	885	670	3	673	78	13	76
Total	1731	470	2201	1364	126	1490	80	27	67

What is the nature of the “unusual episode” to which this table refers?

By Economic Status and Age

Economic Status	Population Exposed to Risk			Number of Deaths			Deaths per 100 Exposed to Risk		
	Adult	Child	Both	Adult	Child	Both	Adult	Child	Both
I (high)	319	6	325	122	0	122	38	0	37
II	261	24	285	167	0	167	64	0	59
III	627	79	706	476	52	528	76	66	73
Other	885	0	885	673	0	673	76	–	76
Total	2092	109	2201	1438	52	1490	69	48	67

Table 2: Population at Risk, Deaths, and Death Rates for the Sinkin_

TU/e

By Economic Status and Sex

Economic Status	Population Exposed to Risk			Number of Deaths			Deaths per 100 Exposed to Risk		
	Male	Female	Both	Male	Female	Both	Male	Female	Both
I (high)	180	145	325	118	4	122	65	3	37
II	179	106	285	154	13	167	87	12	59
III	510	196	706	422	106	528	83	54	73
Other	862	23	885	670	3	673	78	13	76
Total	1731	470	2201	1364	126	1490	80	27	67

What is the nature of the “unusual episode” to which this table refers?

By Economic Status and Age

Economic Status	Population Exposed to Risk			Number of Deaths			Deaths per 100 Exposed to Risk
	Adult	Child	Both	Adult	Child	Both	Adult
I (high)	319	6	325	122	0	122	38
II	261	24	285	167	0	167	64
III	627	79	706	476	52	528	76
Other	885	0	885	673	0	673	76
Total	2092	109	2201	1438	52	1490	69

Table 2: Population at Risk, Deaths, and Death Rates for the Sinkin_

Different columns contain measurements of the same variable: easier to read and interpret but difficult to add data (columns) to the records (rows)

	Class	Sex	Age	Survived	Freq
1	1st	Male	Child	No	0
2	2nd	Male	Child	No	0
3	3rd	Male	Child	No	35
4	Crew	Male	Child	No	0
5	1st	Female	Child	No	0
6	2nd	Female	Child	No	0
7	3rd	Female	Child	No	17
8	Crew	Female	Child	No	0
9	1st	Male	Adult	No	118
10	2nd	Male	Adult	No	154
11	3rd	Male	Adult	No	387
12	Crew	Male	Adult	No	670
13	1st	Female	Adult	No	4
14	2nd	Female	Adult	No	13
15	3rd	Female	Adult	No	89
16	Crew	Female	Adult	No	3
17	1st	Male	Child	Yes	5
18	2nd	Male	Child	Yes	11
19	3rd	Male	Child	Yes	13
20	Crew	Male	Child	Yes	0
21	1st	Female	Child	Yes	1
22	2nd	Female	Child	Yes	13
23	3rd	Female	Child	Yes	14
24	Crew	Female	Child	Yes	0
25	1st	Male	Adult	Yes	57
26	2nd	Male	Adult	Yes	14
27	3rd	Male	Adult	Yes	75
28	Crew	Male	Adult	Yes	192
29	1st	Female	Adult	Yes	140
30	2nd	Female	Adult	Yes	80
31	3rd	Female	Adult	Yes	76
32	Crew	Female	Adult	Yes	20

The same data in a tidy structure (variables in columns and observations in rows)

"The problem is that people like to view data in a totally different way than a computer likes to process it." (Kien Leong)

Tools for tidying data

OpenRefine

- download OpenRefine: http://openrefine.org/download.html
- runs on your computer (not in the cloud), inside the Firefox browser (not in IE), no web connection is needed
- captures all steps done to your raw data ; original dataset is not modified; steps are easily reversed
- with RDF and WikiData extension (FAIRifier)

R, TidyR package

- scripted language (R (free), Matlab, SAS…) to process data (tidying, cleaning, etc.), run the analysis and to produce final outputs

versus

- Excel: data provenance and documentation of data processing with a graphical user interface is bad because it doesn't leaves a record

PROOF course Practical data analysis using R for researchers

See also 'R for Data Science', chapter 12

TU/e

2.b. Making your data understandable for humans #1

Documentation of the table or dataset itself

- columns: use clear, descriptive variable names (no hard to understand abbreviations), avoid special characters (can cause problems with some software)
- rows: if possible, use standard names for nominal/categorical data within cells (derived from a taxonomy, for example: standard species name, CAS registry number for chemical substances…). Use standard date formats.
- try to avoid coding nominal/categorical or ordinal data as numbers
- missing data: use NA

Making your data understandable for humans #2

Documentation of the table or dataset as a whole

The table or dataset contains a description (documentation) that at least mentions:

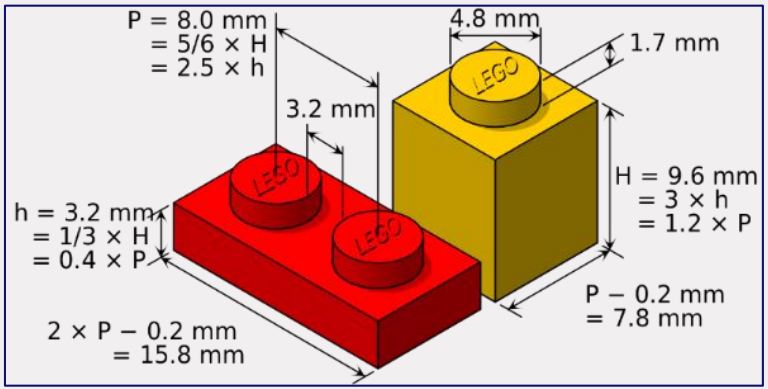
- size of the data set: number of observations and variables
- information about the variables and its measurement units (code book)
- what's included and excluded in the data set, why data are missing
- description of how you collected the data (study design), data manipulation steps (provenance)
- when your data consists of multiple files organized in a folder structure, an explanation of the structure and naming of the files

"Research outputs that are poorly documented are like canned goods with the label removed (…)" (Carly Strasser)

Making your data understandable for humans #3

Metadata standards

Sometimes there are metadata *standards* (here, here) for the documentation or description of your data set but where no standard exists, a simple readme file can be good enough



Making your data understandable for humans #4

- Raw data: https://www.amstat.org/publications/jse/datasets/titanic.dat.txt
- Documentation accompanying the data: https://www.amstat.org/publications/jse/datasets/titanic.txt
- Based on: The "Unusual Episode" Data Revisited / by Robert J. MacG. Dawson, in: *Journal of Statistics Education* vol. 3(1995), issue 3

But sometimes data sets are so complex that a readme file is insufficient

```
NAME:  Population at Risk and Death Rates for an Unusual Episode
TYPE:  Complete record for all of population at risk
SIZE:  2201 observations, 4 variables

DESCRIPTIVE ABSTRACT:
For each person on board the fatal maiden voyage of the ocean liner
Titanic, this dataset records sex, age [adult/child], economic status
[first/second/third class, or crew] and whether or not that person
survived.

SOURCE:
"Report on the Loss of the `Titanic' (S.S.)" (1990), _British Board of
Trade Inquiry Report_ (reprint), Gloucester, UK: Allan Sutton
Publishing.

VARIABLE DESCRIPTIONS:
Column
   1     Class (0 = crew, 1 = first, 2 = second, 3 = third)
  10     Age   (1 = adult, 0 = child)
  19     Sex   (1 = male, 0 = female)
  28     Survived (1 = yes, 0 = no)

Values are aligned and delimited by blanks.  There are no missing
values.

SPECIAL NOTES:
There is not complete agreement among primary sources as to the exact
numbers on board, rescued, or lost.

STORY BEHIND THE DATA:
The sinking of the Titanic is a famous event, and new books are still
being published about it.  Many well-known facts--from the proportions
of first-class passengers to the "women and children first" policy, and
the fact that that policy was not entirely successful in saving the
women and children in the third class--are reflected in the survival
rates for various classes of passenger.  These data were originally
collected by the British Board of Trade in their investigation of the
sinking.
```


TU/e

Making your data findable for humans and search engines

By adding descriptive metadata

- creator
- title
- short description + key words
- date(s) of data collection
- publication year
- related publications
- DOI
- etc.

When uploading your data in a data archive like 4TU.ResearchData, you will be asked to enter these metadata

A DOI is assigned by the data archive

2.c. User license

Make clear *in advance* what other people under *what conditions* are allowed to do with your data by attaching a user license to it

- Creative Commons license for data sets
- GNU General Public License (GPL) for software.
  TU/e example: https://aethelraed.nl/calciumimaginganalyser/index.html)
- License selector ; Choose an open source license

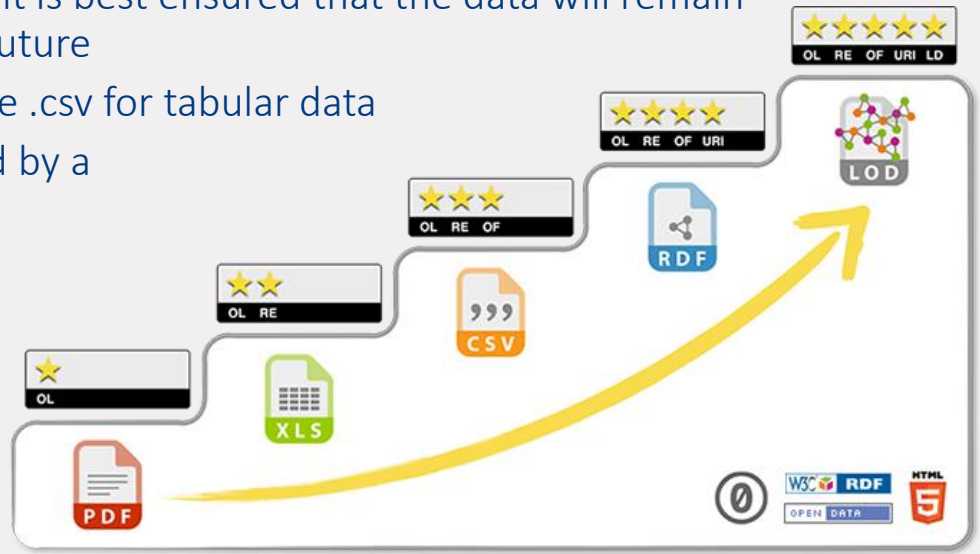
When uploading your data in a data archive like 4TU.ResearchData, you can select a user license of your choice

2.d. Open data formats

Ensuring the ‘longevity’ of your data

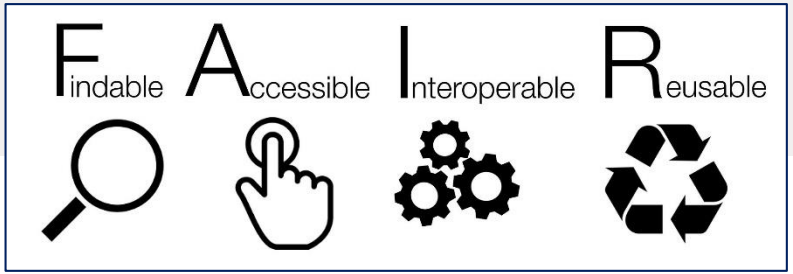
- with open (non-proprietary) data formats it is best ensured that the data will remain usable and ‘legible’ for computers in the future
- are easy to use in a variety of software, like .csv for tabular data
- check the data formats that are supported by a data archive like 4TU.ResearchData

https://5stardata.info/en/

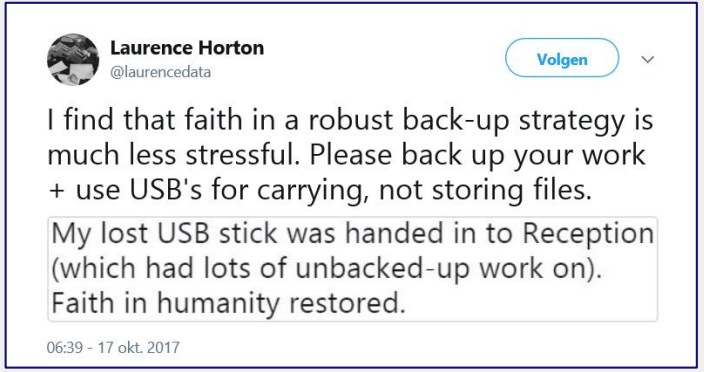


RDM part 3a: Protecting and organizing your data

1. Research data management [RDM]: what and why
2. Caring for your data, or making your data (Re-)usable and Interoperable
    a. tidy data
    b. metadata/documentation
    c. licenses
    d. open data formats
3. Sharing your data, or making your data Findable and Accessible
    a. data protection: back up, file naming, organizing data
    b. data sharing: collaboration platforms, data archives



Protecting your data

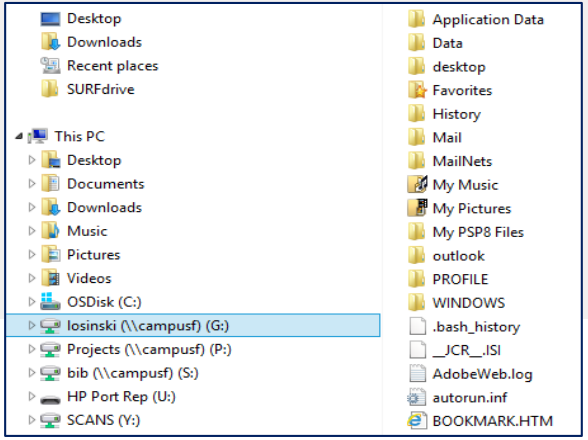


Be safe

1. storage, backup → data safety, protecting against loss
   - 3-2-1 rule: save 3 copies of your data, on 2 different devices, and 1 copy off site
   - use local ICT infrastructure (university network servers: home drives, group drives) if possible

2. access control → data security, protecting against unauthorized use. University network servers are secure but don't allow you to manage access to it.

   With SURFdrive, Open Science Framework or DataverseNL you can manage access to your data yourself



Protecting your data

Be organized

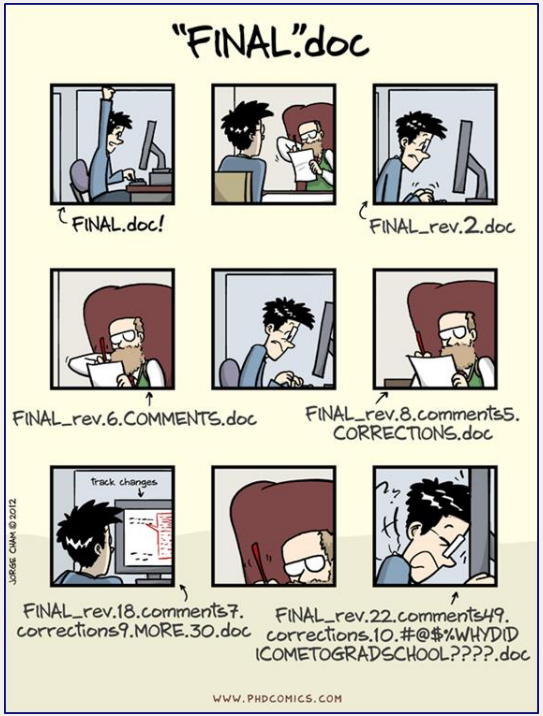
you (and others) should be able to tell what's in a file without opening it

- file-naming
- organizing data in folders

"...we can copy everything and do not manage it well." (Indra Sihar)

TU/e

File naming



Good file names are human readable:

- *meaningful* (use descriptive names that contain info on content)
- *consistent* (use file-naming conventions)
- *unique* (distinguishes a file from files with similar subjects as well as different versions of the file)

and machine readable/searchable

- avoid using special characters in file names
- use "_" underscore to delimit units in names
- use "- " hyphen to delimit names for readability
- include dates (format YYYYMMDD) and a version number on file names

Source: Best practices for file naming (Stanford University Libraries) and http://www2.stat.duke.edu/~rcs46/lectures_2015/01-markdown-git/slides/naming-slides/naming-slides.pdf

Ordering of elements in a file name

Order by date:

2013-04-12_interview-recording_THD.mp3
2013-04-12_interview-transcript_THD.docx
2012-12-15_interview-recording_MBD.mp3
2012-12-15_interview-transcript_MBD.docx

Order by type:

Interview-recording_MBD_2012-12-15.mp3
Interview-recording_THD_2013-04-12.mp3
Interview-transcript_MBD_2012-12-15.docx
Interview-transcript_THD_2013-04-12.docx

Order by subject:

MBD_interview-recording_2012-12-15.mp3
MBD_interview-transcript_2012-12-15.docx
THD_interview-recording_2013-04-12.mp3
THD_interview-transcript_2013-04-12.docx

Forced order with numbering:

01_THD_interview-recording_2013-04-12.mp3
02_THD_interview-transcript_2013-04-12.docx
03_MBD_interview-recording_2012-12-15.mp3
04_MBD_interview-transcript_2012-12-15.docx

Organizing your data in folders

TIER documentation protocol : guiding principles

1. keep your raw or original data raw
   - save your raw data *read-only* in its *original* format in a *separate* folder
   - make a working copy of your raw data (input data, used for processing and analysis)
2. keep the command files (files containing code written in the syntax of the (statistical) software you use for the study) apart from the data
3. keep the analysis files (the fully cleaned and processed data files that you use to generate the results reported in your paper) in a separate folder
4. store the metadata (codebook, description of variables, etc.) in a separate folder, apart from the data itself

Organizing your data in folders

1. Main project folder (name of your research project/working title of your paper)
    1.1. Original data and metadata
        1.1.1. Original data
        1.1.2. Metadata
    1.2. Processing and analysis files
        1.2.1. Importable data files
        1.2.2. Command files
        1.2.3. Final data files
    1.3. Documents
    1.4. Literature

Organizing your data in folders

1. Main project folder (name of your research project/working title of your paper)
    1.1. Original data and metadata
        1.1.1. Original data (raw data, obtained/gathered data)
        - Any data that were necessary for any part of the processing and/or analysis you reported in your paper.
        - Copies of all your original data files, saved in exactly the format it was when you first obtained it. The name of the original data file may be changed.
        - Keep these data read only!

        1.1.2. Metadata

Organizing your data in folders

1. Main project folder (name of your research project/working title of your paper)
    1.1. Original data and metadata
        1.1.1. Original data
        1.1.2. Metadata (applies to obtained data files)
        - The Metadata Guide: document that provides information about each of your original data files that is not written by yourself but that is written in existing supplementary documents, such as users' guides and code books that accompany the original data file
        - A bibliographic citation of the original data files, including the date you downloaded or obtained the original data files and unique identifiers that have been assigned to the original data files.
        - Information about how to obtain a copy of the original data file
        - Whatever additional information to understand and use the data in the original data file

Organizing your data in folders

1.1. Original data and metadata

1.2. Processing and analysis files

1.2.1. Importable data files (the data you work with, input data, suitable for processing and analysis)

- A corresponding version for each of the original data files. This version can be identical to the original version, or in some cases it will be a modified version. For example modifications required to allow your software to read the file (converting the file to another format, removing unusable data or explanatory notes from a table)
- The original and importable versions of a data file should be given different names
- The importable data file should be as nearly as identical as possible to the original
- The changes you make to your original data files to create the corresponding importable data files should be described in a Readme file

1.2.2. Command files

1.2.3. Final data files

Organizing your data in folders

1.1. Original data and metadata

1.2. Processing and analysis files

1.2.1. Importable data files

1.2.2. Command files

One or more files containing code written in the syntax of the (statistical) software you use for the study

- Importing phase: commands to import or read the files and save them in a format that suits your software
- Processing phase: commands that execute all the processing required to transform the importable version of your files into the final data files that you will use in your analysis (i.e. cleaning, recoding, joining two or more data files, dropping variables or cases, generating new variables)
- Generating the results: commands that open the final data file(s), and then generate the results reported in your paper.

1.2.3. Final data files

Organizing your data in folders

1.1. Original data and metadata

1.2. Processing and analysis files

1.2.1. Importable data files

1.2.2. Command files

1.2.3. Final data files

- The fully cleaned and processed data files that you use to generate the results reported in your paper
- The Data Appendix: codebook for your final data files: brief description of the analysis data file(s), a complete definition of each variable (including coding and/or units of measurement), the name of the original data files from which the variable was extracted, the number of valid observations for the variable, and the number of cases with missing values

Organizing your data in folders

1.1. Original data and metadata

1.2. Processing and analysis files

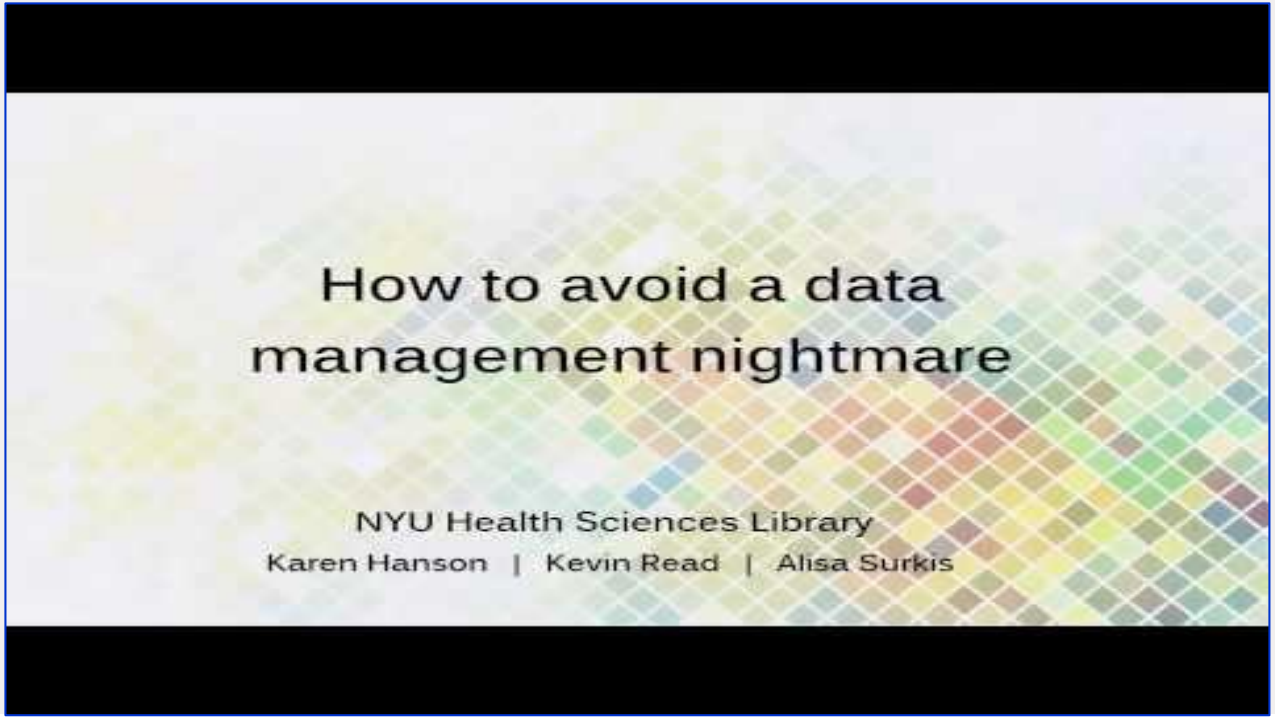
1.3. Documents

- An electronic copy of your complete final paper
- The Readme-file for your replication documentation
- What statistical software or other computer programs are needed to run the command files
- Explanation of the structure or hierarchy of folders in which the data is stored
- Describe precisely any changes you made to your original data files to create the corresponding importable data files
- Step-by-step instructions for using your documentation to replicate the statistical results reported in your paper

1.4. Literature

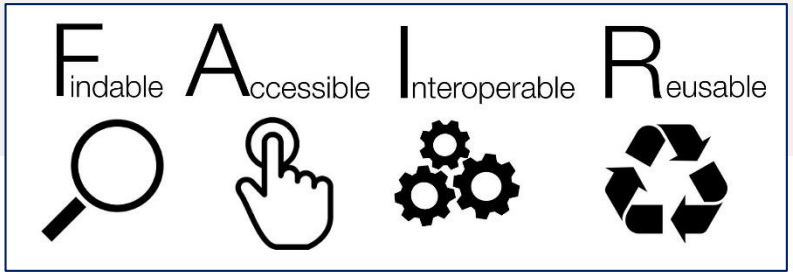
- Retrieved relevant literature

Summary so far



RDM part 3b: data sharing

1. Research data management [RDM]: what and why
2. Caring for your data, or making your data (Re-)usable and Interoperable
   a. tidy data
   b. metadata/documentation
   c. licenses
   d. open data formats
3. Sharing your data, or making your data Findable and Accessible
   a. data protection: back up, file naming, organizing data
   b. data sharing: collaboration platforms, data archives



Why sharing research data?

- Because you work together with other researchers → *collaborative science*
- Because of re-using results → *data-driven science*, *open science*
- Because of scientific integrity: validating data analysis by reproducibility checks requires data and the code that is used to clean, process and analyze the data and to produce the final outputs → *reproducible science*

Additional reasons

- Because your data are unique / not easily repeatable (long term observational data)
- Because it's required… by journals, by funders like NWO and EC and by your university

Why sharing research data?

TUE code of scientific conduct

3. Openness

Open and unbiased communication is essential for science and engineering. For academic staff and students, this entails that:

- They contribute actively to an academic climate in which insights and criticisms are welcome from all, regardless of academic rank and personal characteristics.
- They give room to others to develop or take their own intellectual stance in research, design and education.
- Whenever they publish research results, they present their research such that its results may in principle be replicated.
- They make accessible, after publication, all information needed for intersubjective testing of design results and design processes.
- They make accessible, after publication, research data for re-use by colleagues.

Sharing your data

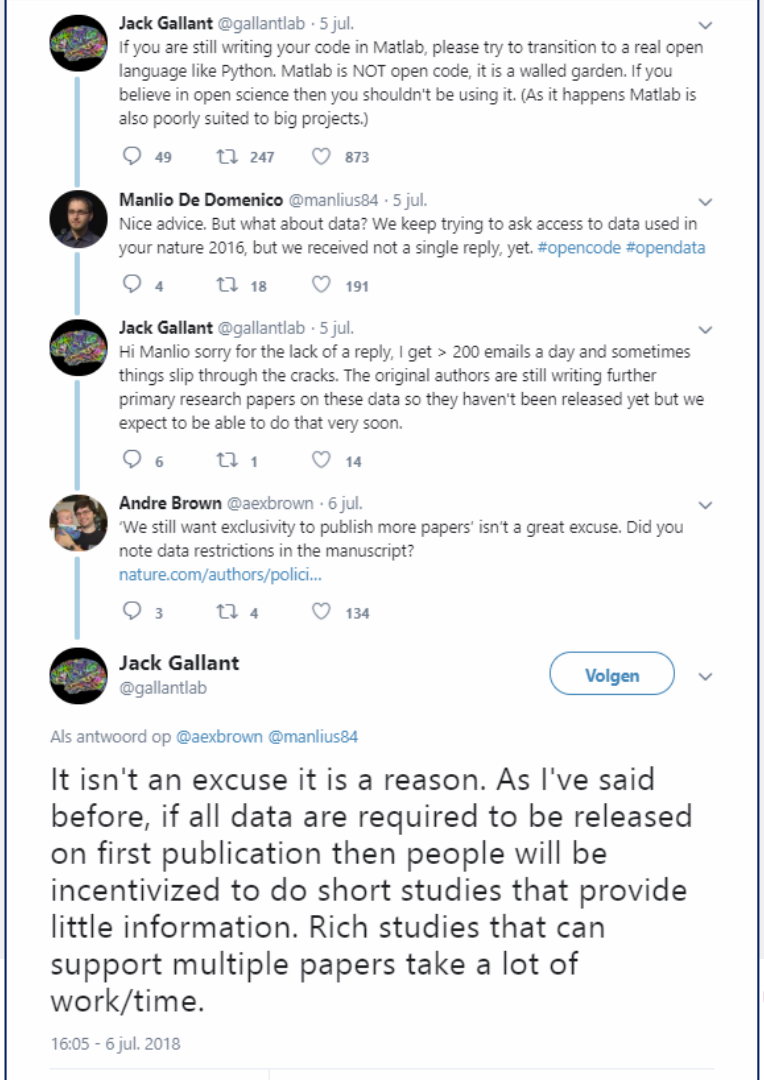
Why don’t people always do it?

When you approach a PI for the source codes and raw data, you better explain who you are, whom you work for, why you need the data and what you are going to do with it.

I have to say that this is a very unusual request without any explanation! Please ask your supervisor to send me an email with a detailed, and I mean detailed, explanation.

The data files remains our property and are not deposited for free access. Please, let me know the purpose you want to get the file and we will see how we can help you.

We do not typically share our internal data or code with people outside our collaboration.



Reasons not to share your data

- Preparing my data for sharing takes time and effort
  - But research data management also increases your research efficiency
- My data are confidential
  - But you can anonymize or pseudonymize your data
- My data still need to yield publications
  - But you can publish your data under an embargo and by publishing your data you establish priority and you can get credits for it
- My data can be misused or misinterpret
  - But the best defense against malicious use is to refer to an archival copy of your data which is guaranteed exactly as you mean it to be
- My data are only interesting for me
  - But sharing your data may be required by a funder / journal or your data may be requested to validate your results

Sharing your data

During your research versus after your research

During your research via collaboration or sharing platforms

- Data sharing is (more) aimed at collaboration, at working together on data
- Being able to control access to data is crucial

After your research via data archives or repositories

- Keeping and, if necessary, publishing an archival copy of data – a copy that cannot be changed - is essential
- Long term preservation of your data - at least 10 years - is important

Sharing data during your research

General data sharing tools

- SURFdrive [TU/e only]: Dutch academic Dropbox, 250 Gb, maximum data transfer 16 Gb
- OneDrive (supported by TUE)
- Google Drive, Dropbox: don't use these to store sensitive data

Sharing data during your research

DataverseNL [TU/e only]: data sharing platform for *active* research data [based on Harvard's Dataverse Project] where you may:

- store your data in an organized and safe way
- clearly describe your data
- version control of your data
- arrange access to your data

If you are interested in using DataverseNL, please contact me (l.osinski@tue.nl)

You may use DataverseNL

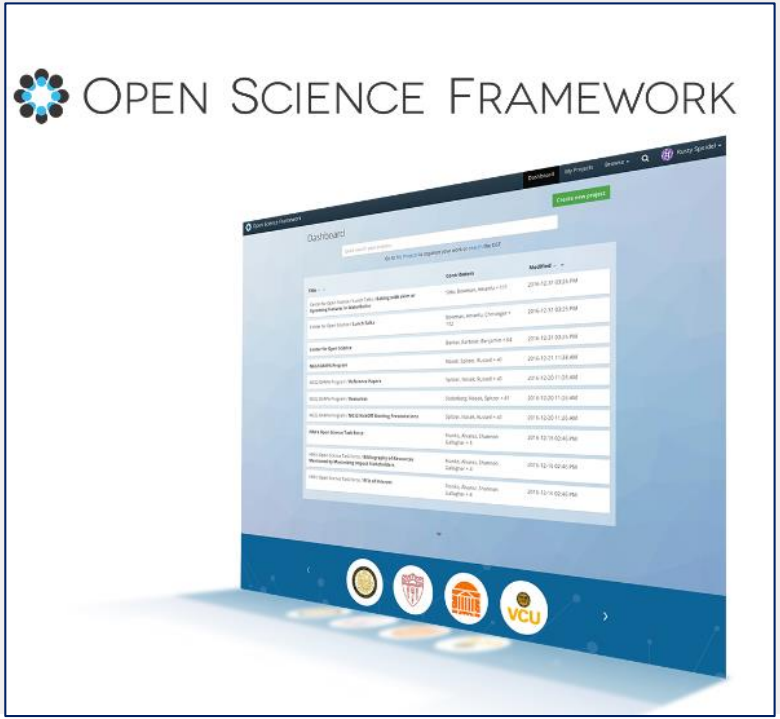
- Go to: https://dataverse.nl/ or https://act.dataverse.nl (demo version)
- Click 'Log in' (at the top right)
- Click 'Institutional login'
- First time: select Eindhoven University of Technology and log on with your TU/e username and password
- First time: when asked for it, give permission to share your data by answering Yes or click this Tab
- First time: when asked to create an account, answer Yes or click this Tab.
- When you succeeded to create an account, your username is the prefix of your email address

Click 4TU dataverse → Eindhoven dataverse → Add data: you can now create and publish data sets, upload files and assign access rights to data sets or files.

Sharing data during your research

Open Science Framework

- Peder Isager, How to share your data online with OSF
- Courtney Soderberg, Using OSF to share data: a step by step guide



Sharing data after your research

On request

"I'd like to thank E.J. Masicampo and Daniel LaLande for sharing and allowing me to share their data…"

Daniël Lakens (2014), What p-hacking really looks like: A comment on Masicampo & LaLande (2012)

On a (personal) website

"Let me start by saying that the reason why I put all excel files online, including all the detailed excel formulas about data constructions and adjustments, is precisely because I want to promote an open and transparent debate about these important and sensitive measurement issues."

Thomas Piketty, My response to the Financial Times, HuffPost The Blog, 29-05-2014 ; originally published as Addendum: Response to FT, 28-05-2014

A data journal

Journal of open psychology data, Geoscience data journal, Data in brief, Scientific data

Sharing data after your research

Via an archive or repository

Choose a repository where other researchers in your discipline are sharing their data, for example TurBase (turbulence data), Lxcat (plasma data).

If not available, use a multidisciplinary or general repository that at least assigns a persistent identifier to your data (DOI) and requires that you provide adequate metadata
for example Zenodo, Figshare, DANS or:

4TU.ResearchData

4TU.Centre for Research Data is for the *publication* of static data ('frozen' data sets, 'milestone' data sets) after the project has ended.

You can upload your data yourself (single data sets < 3Gb)

4TU.Centre for Research Data and FAIR

With 4TU.ResearchData data are made FAIR to a certain extent

- Data are assigned a DOI
- Data can be linked to publications (DOI reservation is possible)
- Data are assigned descriptive/discovery metadata
- Data are assigned a user license of choice
- Data are open access (restricted access options being developed)
- Data are archived/preserved for the long term
- Metadata can be harvested by Google etc.

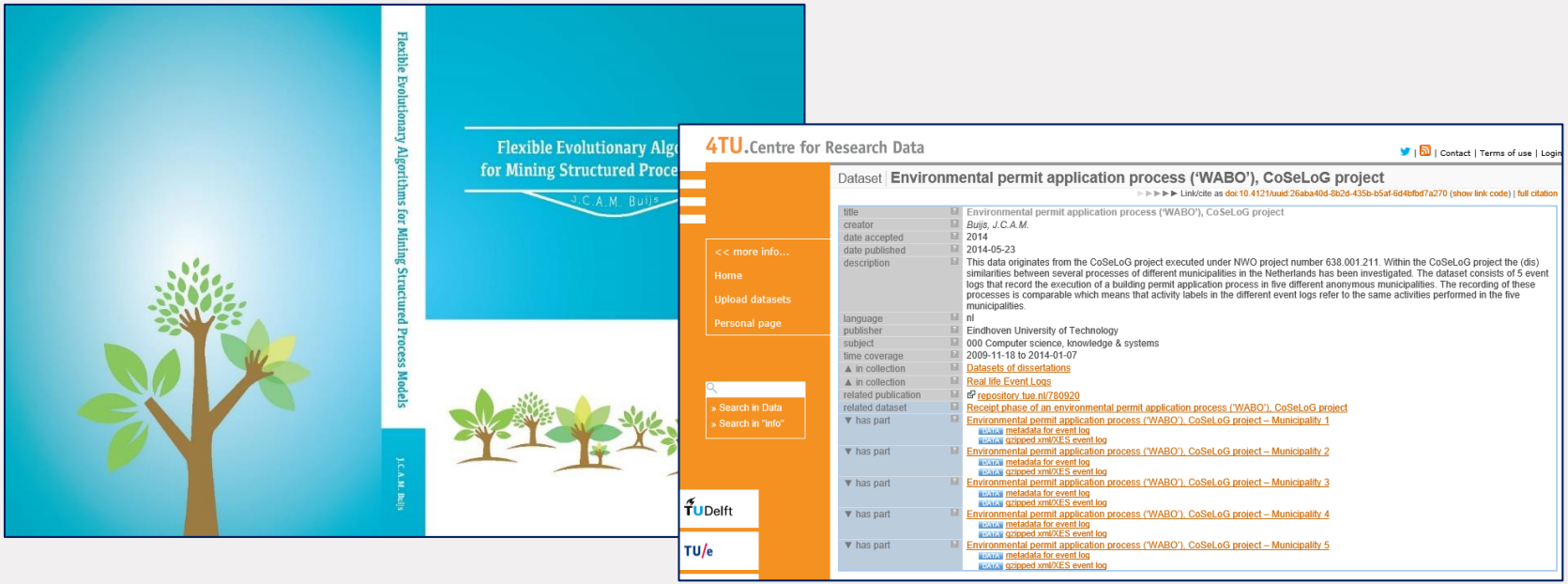
4TU.Centre for Research Data and FAIR

With 4TU.ResearchData data are made FAIR to a certain extent

- Data are assigned a DOI → findable
- Data can be linked to publications (DOI reservation is possible) → findable
- Data are assigned descriptive/discovery metadata → findable, interoperable
- Data are assigned a user license of choice → re-useable
- Data are open access (restricted access options being developed) → accessible
- Data are archived/preserved for the long term → accessible
- Metadata can be harvested by Google etc. → findable

Sharing data after your research

Link your data to your publication



Recommended reading on ‘good data practices’

1. Goodman, A., Pepe, A., Blocker, A.W., et al. (2014) Ten simple rules for the care and feeding of scientific data, *PLOS Computional Biology, 10*(4), e10033542. https://doi.org/10.1371/journal.pcbi.1003542
2. Eugene Barsky (2017), Good enough research data management: a very brief guide
3. Broman, K.W., Woo, K.H., Data organization in spreadsheets, in: The American Statistician, https://doi.org/10.1080/00031305.2017.1375989

---

4. Ellis SE, Leek JT. (2017) How to share data for collaboration. *PeerJ reprints* : e3139v1 https://doi.org/10.7287/peerj.preprints.3139v1
5. Dynamic ecology (2016), Ten commandments for good data management. https://dynamicecology.wordpress.com/2016/08/22/ten-commandments-for-good-data-management/
6. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. *PLOS Computational Biology, 13*(6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510

More info

- Data Coach [website]
- Working with data [website]
- Online course with web lectures

Mentioned URL's (in order of appearance)

What and why

1. Figshare support, The importance of data management for research: https://youtu.be/Ae205CNrk6w
2. Henry Rzepa, Collaborative FAIR data sharing: http://www.ch.imperial.ac.uk/rzepa/blog/?p=16292
3. The FAIR guiding principles for scientific datamanagement and stewardship: https://doi.org/10.1038/sdata.2016.18

(Re-)usable data

4. Hashtag #otherpeoplesdata (twitter): https://twitter.com/hashtag/otherpeoplesdata
5. Bad data: http://okfnlabs.org/bad-data/
6. Bad data Nature magazine: http://okfnlabs.org/bad-data/ex/nature-magazine-supplementary/
7. Nature article: Humayun, M., e.a., Origin and age of the earliest Martian crust from meteorite NWA 7533. https://dx.doi.org/10.1038/nature12764
8. Supplematary information belonging to Nature article: https://media.nature.com/original/nature-assets/nature/journal/v503/n7477/extref/nature12764-s1.pdf
9. Tidy data: https://www.jstatsoft.org/article/view/v059i10
10. OpenRefine: http://openrefine.org
11. TidyR: http://tidyr.tidyverse.org/
12. R: https://www.r-project.org/

Mentioned URL's (in order of appearance)

13. PROOF course Practical data analysis using R for researchers: https://intranet.tue.nl/en/university/services/service-for-personnel-and-organization/human-resource-management/professional-development/proof-training-program/research-skills/practical-data-analysis-using-r-for-researchers/
14. CAS registry number: https://www.cas.org/support/documentation/chemical-substances
15. RDA metadata directory: http://rd-alliance.github.io/metadata-directory/
16. Metadata standards: https://fairsharing.org
17. Readme file: https://researchdata.4tu.nl/fileadmin/editor_upload/pdf/README/Guidelines_for_creating_a_README_file.pdf
18. Uploading your data with 4TU.ResearchData: https://researchdata.4tu.nl/en/use-4turesearchdata/archive-research-data/
19. Licensing your data with 4TU.ResearchData: https://researchdata.4tu.nl/en/use-4turesearchdata/archive-research-data/upload-your-data-in-our-data-archive/licencing/
20. Creative Commons licenses: https://creativecommons.org/
21. GNU General Public License: https://www.gnu.org/licenses/gpl-3.0.en.html
22. TU/e example GPL license: https://aethelraed.nl/calciumimaginganalyser/index.html
23. License selector: https://ufal.github.io/public-license-selector/
24. Tim Berners Lee, 5 star open data: https://5stardata.info/en/
25. Preferred data formats of 4TU.ResearchData: http://researchdata.4tu.nl/en/publishing-research/data-description-and-formats/

Mentioned URL's (in order of appearance)

Protecting your data

26. Storage, back up of data: https://www.ukdataservice.ac.uk/manage-data/store
27. Local ICT infrastructure: https://intranet.tue.nl/en/university/services/ict-services/ict-service-catalog/management-services/data-management-storage/ (TU/e intranet)
28. SURFdrive (at TU/e): https://intranet.tue.nl/en/university/services/ict-services/ict-service-catalog/management-services/data-management-surfdrive
29. Open Science Framework: https://osf.io/
30. DataverseNL: https://dataverse.nl/dvn/
31. "Final".doc (cartoon): http://phdcomics.com/comics/archive.php?comicid=1531
32. Best practices for file naming: http://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming
33. Jenny Bryan, Naming things: http://www2.stat.duke.edu/~rcs46/lectures_2015/01-markdown-git/slides/naming-slides/naming-slides.pdf
34. Project TIER: https://www.projecttier.org/
35. TIER documentation protocol: https://www.projecttier.org/tier-protocol/specifications/#overview-of-the-documentation
36. How to avoid a data management nightmare: https://youtu.be/nNBiCcBlwRA

Mentioned URL's (in order of appearance)

Sharing your data

37. Olivier H. Beauchesne, Map of scientific collaborations (Redux): http://olihb.com/2014/08/11/map-of-scientific-collaboration-redux/
38. Victoria Stodden, Jennifer Seiler, and Zhaokun Ma, An empirical analysis of journal policy effectiveness for computational reproducibility. https://doi.org/10.1073/pnas.1708290115
39. NWO and research data: http://www.nwo.nl/en/policies/open+science/data+management
40. Horizon 2020 Guidelines on data management: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
41. TU/e Code of Scientific Conduct: https://www.tue.nl/en/our-university/about-the-university/organization/integrity/scientific-integrity/
42. Why don't people always share data (twitter): https://twitter.com/cj_battey/status/974045242444820480
43. Data sharing discussion (twitter): https://twitter.com/gallantlab/status/1015371268386770945
44. Emilio M. Bruna (04-09-2014), The opportunity cost of my #OpenScience was 36 hours + $690 (UPDATED) . http://brunalab.org/blog/2014/09/04/the-opportunity-cost-of-my-openscience-was-35-hours-690/
45. Rouder, Jeffrey N., The what, why, and how of born-open data, Behavior Research Methods, vol. 48(2016), p. 1062-1069. http://dx.doi.org/10.3758/s13428-015-0630-z (see p. 1063: "It was a pain to document the data; it was a pain to format the data")
46. Amnesia, data anonymization tool: https://amnesia.openaire.eu/index.html
47. SURFdrive: https://www.surfdrive.nl/
48. OneDrive: https://intranet.tue.nl/en/university/services/01-01-1970-information-management-services/help-and-support/manuals/user-support-systems/office-365/manual-transition-to-office-365/office-365-parts-and-items/onedrive/

Mentioned URL's (in order of appearance)

49. Google Drive: https://www.google.com/drive/
50. Dropbox: https://www.dropbox.com/
51. Peder Isager, How to share your data online with OSF: https://pedermisager.netlify.com/post/how-to-share-your-data-with-osf/
52. Courtney Soderberg, Using OSF to share data: a step by step guide: https://doi.org/10.1177%2F2515245918757689
53. Data on request (blog post Daniel Lakens): http://daniellakens.blogspot.nl/2014/09/what-p-hacking-really-looks-like.html
54. Data on personal website (Thomas Piketty): http://piketty.pse.ens.fr/en/capital21c2
55. Journal of open psychology data: https://openpsychologydata.metajnl.com/
56. Geoscience data journal : http://rmets.onlinelibrary.wiley.com/hub/journal/10.1002/(ISSN)2049-6060/
57. Data in brief : https://www.journals.elsevier.com/data-in-brief
58. Scientific data: https://www.nature.com/sdata/
59. TurBase: https://turbase.cineca.it/
60. LXcat: https://fr.lxcat.net/home/
61. Research data catalogue Re3data.org: https://www.re3data.org/
62. Publishing data: DANS: http://www.dans.knaw.nl/en
63. Publishing data: 4TU.Centre for Research Data: https://researchdata.4tu.nl/en/

Mentioned URL's (in order of appearance)

64. Publishing data: Zenodo: http://www.zenodo.org/
65. Publishing data: Figshare: http://www.figshare.com
66. Data Seal of Approval https://datasealofapproval.org/en/
67. Self upload 4TU.ResearchData: https://data.4tu.nl/account/login/?next=/upload/
68. Data sets underlying PhD thesis Joos Buijs: http://dx.doi.org/10.4121/uuid:26aba40d-8b2d-435b-b5af-6d4bfbd7a270
69. PhD thesis Joos Buijs: http://dx.doi.org/10.6100/IR780920

Mentioned URL's (in order of appearance)

Recommended reading and more info

70. Goodman, A., Pepe, A., Blocker, A.W., et al. (2014) Ten simple rules for the care and feeding of scientific data, *PLOS Computional Biology, 10*(4), e10033542. https://doi.org/10.1371/journal.pcbi.1003542
71. Eugene Barsky (2017), Good enough research data management: a very brief guide
72. Dynamic ecology (2016), Ten commandments for good data management. https://dynamicecology.wordpress.com/2016/08/22/ten-commandments-for-good-data-management/
73. Ellis SE, Leek JT. (2017) How to share data for collaboration. *PeerJ Preprints*5:e3139v1 https://doi.org/10.7287/peerj.preprints.3139v1
74. Dynamic ecology (2016), Ten commandments for good data management. https://dynamicecology.wordpress.com/2016/08/22/ten-commandments-for-good-data-management/
75. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. *PLOS Computational Biology, 13*(6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510
76. Data Coach (website): https://www.tue.nl/datacoach
77. Working with data (website): https://intranet.tue.nl/en/university/digital-university/data-stewardship/working-with-data/
78. Online course with weblectures: https://intranet.tue.nl/universiteit/diensten/dienst-personeel-en-organisatie/human-resource-development/professional-development/research-and-data-management-rdm/