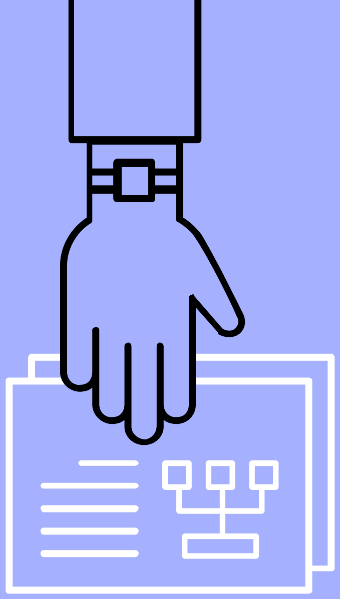
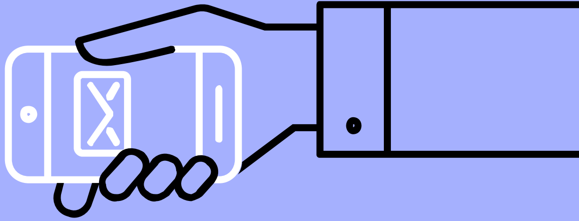
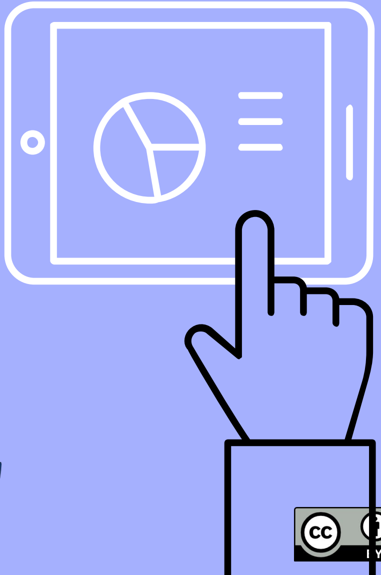


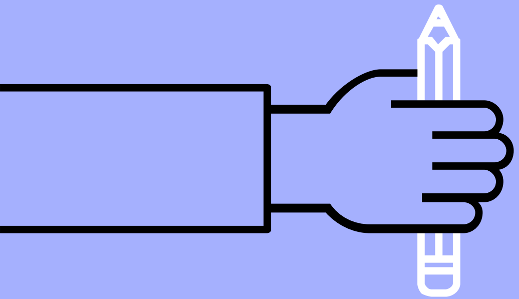


FAIR Data & Data Management Plans

@pcmasuzzo

Eurodoc OS Ambassadors
Webinar - 11/04/2019

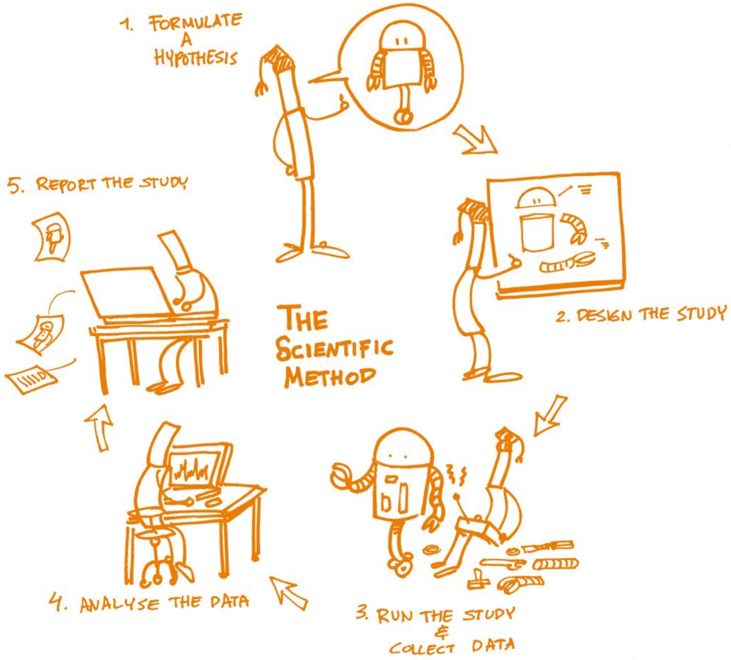


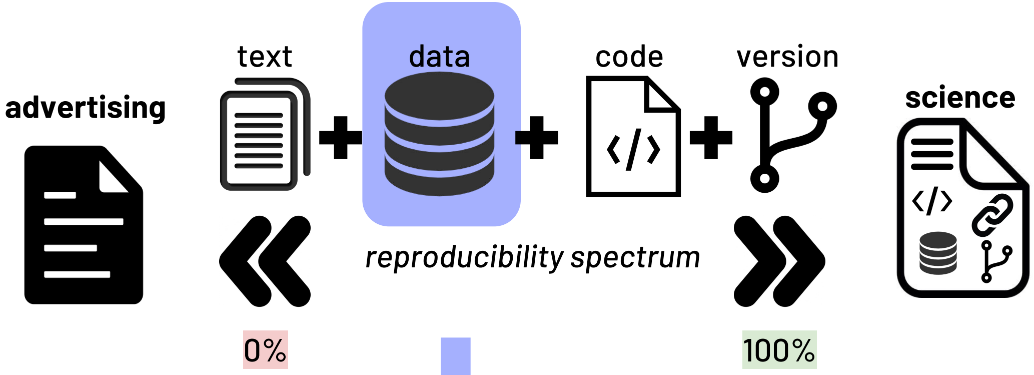


OPEN
SCIENCE
MOOC
FREE | OPEN | LEARNING

IGDORE
The Globally Distributed Institute for
Open Research and Education

Research data are first-class citizens in science




advertising

text + data + code + version

science

« *reproducibility spectrum* »

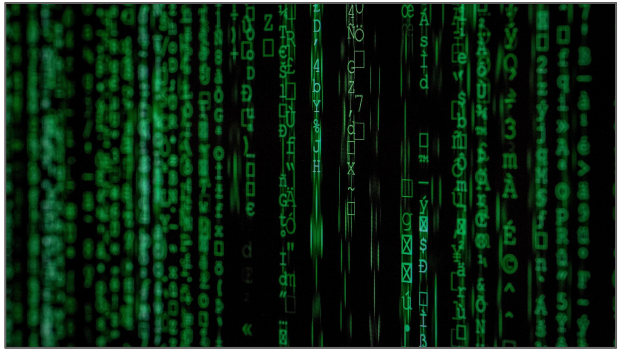
0%

100%

but who does really care about these data, to begin with?

[Open Science Training Handbook](), Foster

The short answer: everybody interested in (modern) science



Researchers AND machines need to find/discover data having features of interest, for which they will be using links, metadata, as well as actual data

Once found, machines need to access/retrieve data of interest (obtain a copy of the contents in some format)

Finally, researchers go to their computers and start to re-use/analyze the data; they need to have workflow tools to:
-aggregate data and run analyses,
-reproduce existing pipelines,
-validate sanity of data,
-....

Unsplash photos

In the eScience ecosystem, enabling optimal use of research data and methods is a big challenge

Researchers ready to share their data and interpretations

Professional data publishers (data repositories, data journals)

Funding agencies increasingly concerned about data stewardship

$$$ ✓

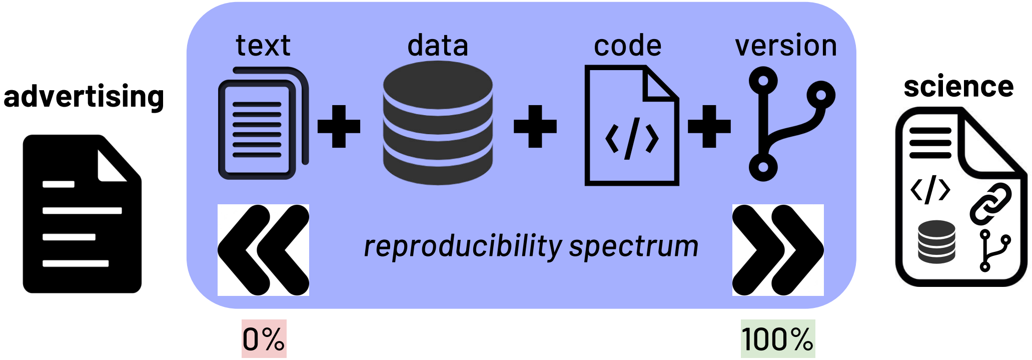
Data science community analysing data to advance discovery

Providing **machine-readable data as the main substrate for Knowledge Discovery** is a big challenge in modern Science

The **best practice recommendation for research data** is to be as open and FAIR as possible
(while accounting for ethical, commercial and privacy constraints with sensitive data or proprietary data)

Force11 FAIR principles

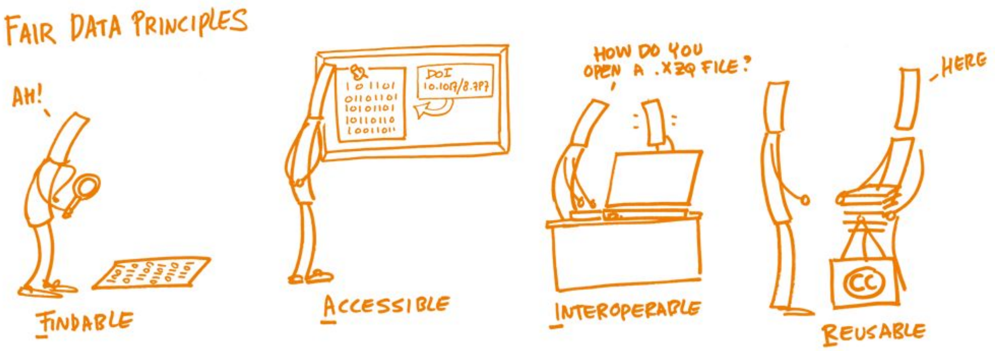
Research objects: scholarly information on the Web



The basic idea is that objects that belong together (*e.g.*, an article with its associated code, data and workflows) should have some means of being aggregated, so that all associated research objects can be discovered together.

Although this might seem to be obvious, as research objects are scattered across different repositories on the web, the connections between them are often lost.

The FAIR principles provide guidance for scientific data management and stewardship



1) FAIR *does not* imply Open (but FAIR data can definitely be Open Data; the two concepts can overlap)
2) the FAIR facets are related to each other but technically somehow independent of each other (we can speak of degrees of FAIR-ness)
3) the FAIR principles are agnostic of technical implementations (FAIR-ness can be achieved with a wide range of technologies and implementations)

the principles describe characteristics that data should exhibit to **assist discovery and reuse through the web**

Findable: data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier

Accessible: metadata and data are understandable to humans and machines and deposited in a trusted repository

Interoperable: metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation

Reusable: data and collections have a clear usage licenses and provide accurate information on provenance

Open Science Training Handbook, Foster; Force11 FAIR principles

Findable

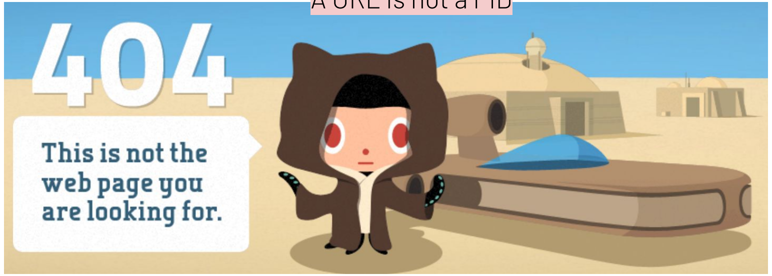
F	Findable
F1	(meta)data are assigned a globally unique and persistent identifier
F2	data are described with rich metadata
F3	metadata clearly and explicitly include the identifier of the data it describes
F4	(meta)data are registered or indexed in a searchable resource

Findable

F	Findable
F1	**(meta)data are assigned a globally unique and persistent identifier**
F2	data are described with rich metadata
F3	metadata clearly and explicitly include the identifier of the data it describes
F4	(meta)data are registered or indexed in a searchable resource

[DOI [Digital Object Identifier]]
the most widely used PID for research data

A URL is not a PID



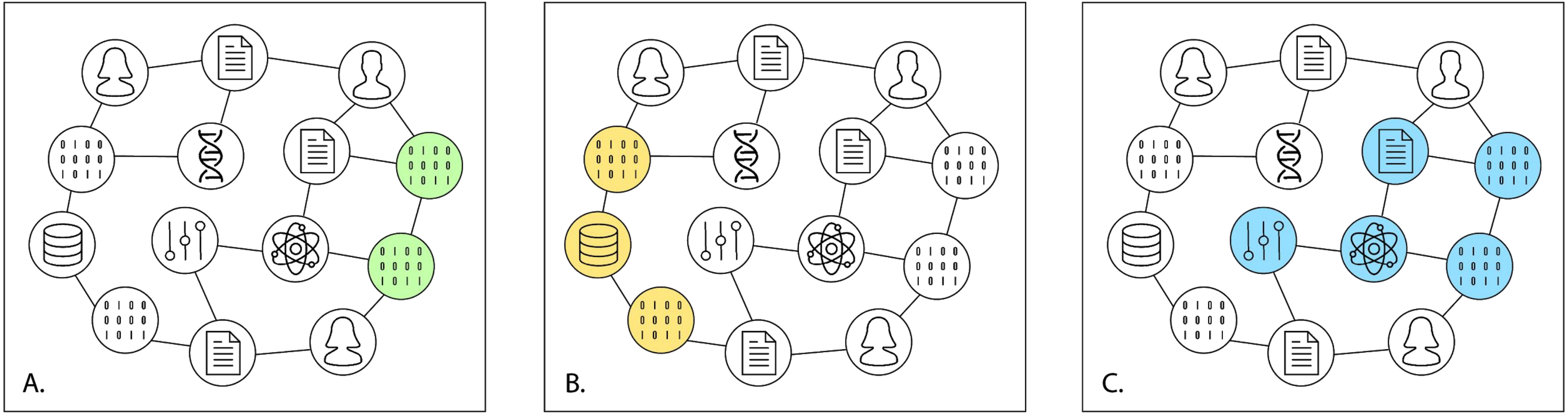
the dataset needs to be identified by a Persistent Identifier [PID]
so that it can be located by a machine

PID = a **globally unique** and **long-lasting** reference to something, *e.g.*, documents, files, books, people

A PID is **separated from location**: if a web document is moved, the PID points to the same object in the new location

Best practice
1. Deposit your data to a domain-specific repository
2. If that does not exist (yet), use general-purpose repositories

Towards a PID graph to enable discovery and impact assessment



Different versions of software code

Datasets hosted by a particular repository

All digital objects connected to a research object

Introducing the PID graph; Research Data Alliance DDRI Working group

Data Sharing or Data Visitation?

Findable where?

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

NADAC
National Archive of Data on Arts & Culture

RCSB PDB
PROTEIN DATA BANK

IDR

NEURO VAULT

CXIDB
Coherent X-ray Imaging Data Bank

figshare

zenodo

OpenfMRI

DRYAD

COS
CENTER FOR OPEN SCIENCE

OpenTrials
All the Data, on All the Trials, Linked

OPEN ICPSR

https://fairsharing.org/databases/, https://www.re3data.org/

Findable

F	Findable
F1	(meta)data are assigned a globally unique and persistent identifier
F2	**data are described with rich metadata**
F3	metadata clearly and explicitly include the identifier of the data it describes
F4	(meta)data are registered or indexed in a searchable resource

the data are described by rich metadata, so they can be discovered by a human

Your first collaborator is yourself, and your past self does not answer emails.

Intrinsic metadata: immutable
the author of a book
the date a photo was taken

Extrinsic metadata: depend on the context
the date a book was purchased
the publications a photo has appeared in

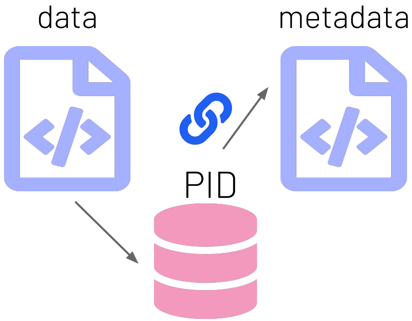
Best practice

1. Never presume that you know who will want to use your data, or for what purpose: be generous!
2. Whenever possible, use community standards
3. Have someone "naive" check your annotations

RDM promotion resources, list of metadata standards, FAIRsharing

Findable

F	Findable
F1	(meta)data are assigned a globally unique and persistent identifier
F2	data are described with rich metadata
F3	**metadata clearly and explicitly include the identifier of the data it describes**
F4	(meta)data are registered or indexed in a searchable resource



the data are described by rich metadata, so they can be discovered by a human

Your first collaborator is yourself, and your past self does not answer emails.

Intrinsic metadata: immutable
the author of a book
the date a photo was taken

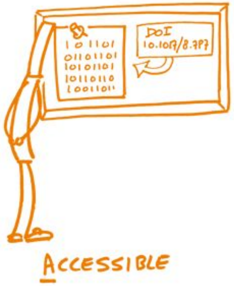
Extrinsic metadata: depend on the context
the date a book was purchased
the publications a photo has appeared in

Best practice
1. Never presume that you know who will want to use your data, or for what purpose: be generous!
2. Whenever possible, use community standards
3. Have someone "naive" check your annotations
4. IF you do F1 correctly, F3 comes for free!

RDM promotion resources, list of metadata standards, FAIRsharing

Accessible

A	Accessible
A1	(meta)data are retrievable by their identifier using a standardized communications protocol
A1.1	the protocol is open, free, and universally implementable
A1.2	the protocol allows for an authentication and authorization procedure, where necessary
A4	metadata are accessible, even when the data are no longer available



Accessible

A	Accessible
A1	**(meta)data are retrievable by their identifier using a standardized communications protocol**
A1.1	**the protocol is open, free, and universally implementable**
A1.2	**the protocol allows for an authentication and authorization procedure, where necessary**
A4	metadata are accessible, even when the data are no longer available

limitations on and protocols for
the use of data are made explicit

Data should be retrievable by anyone with a computer and an internet connection, if they are authorized, with a well-defined protocol.

Accessible data does not automatically imply open or free access: data published with restricted access can also be FAIR.

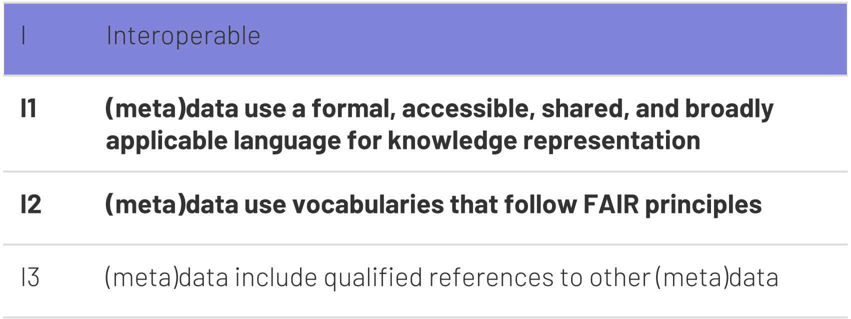
Best practice

1. Include **clear** licenses and conditions of use: who can access the data and in what way?
2. Remember: no license does not mean OK to access
   (when in doubt, the answer is NO)

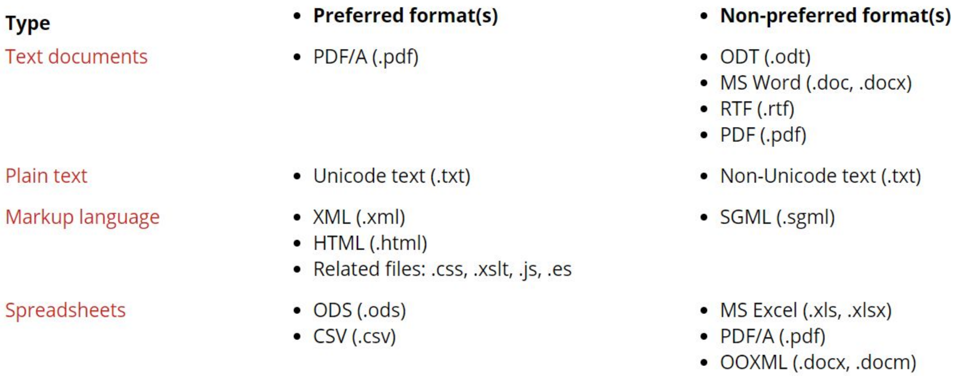
Interoperable

I	Interoperable
I1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2	(meta)data use vocabularies that follow FAIR principles
I3	(meta)data include qualified references to other (meta)data

Interoperable



I	Interoperable
I1	**(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation**
I2	**(meta)data use vocabularies that follow FAIR principles**
I3	(meta)data include qualified references to other (meta)data



Type	**Preferred format(s)**	**Non-preferred format(s)**
Text documents	PDF/A (.pdf)	ODT (.odt) MS Word (.doc, .docx) RTF (.rtf) PDF (.pdf)
Plain text	Unicode text (.txt)	Non-Unicode text (.txt)
Markup language	XML (.xml) HTML (.html) Related files: .css, .xslt, .js, .es	SGML (.sgml)
Spreadsheets	ODS (.ods) CSV (.csv)	MS Excel (.xls, .xlsx) PDF/A (.pdf) OOXML (.docx, .docm)

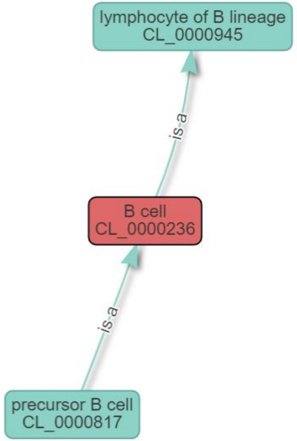
https://en.wikibooks.org/wiki/Choosing_The_Right_File_Format/Quick_Guide

People and computers can interpret your data and combine them with other datasets

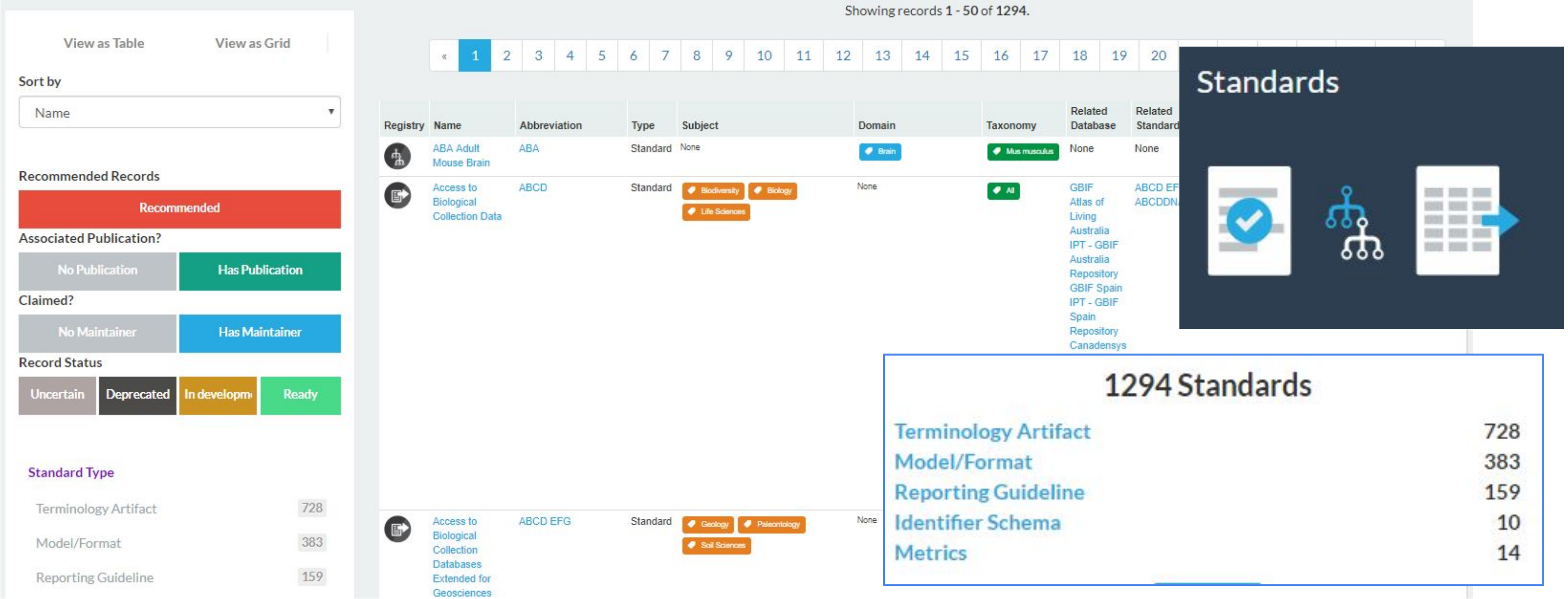
Clearly, this is a very challenging requirement to meet.

Best practice

1. Document as much as possible - be generous!
2. Use *preferred file formats*
3. If existing in your discipline, use standard vocabularies, ontologies and thesauri in your (meta)data, or provide mapping of your data to these vocabularies, ontologies and thesauri



Metadata, Ontologies, and Vocabularies Standards help Achieve Interoperability Principles



View as Table

View as Grid

Sort by

Name

Recommended Records

Recommended

Associated Publication?

No Publication | Has Publication

Claimed?

No Maintainer | Has Maintainer

Record Status

Uncertain | Deprecated | In developm | Ready

Standard Type

Terminology Artifact	728
Model/Format	383
Reporting Guideline	159

Showing records 1 - 50 of 1294.

« 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Registry	Name	Abbreviation	Type	Subject	Domain	Taxonomy	Related Database	Related Standard
	ABA Adult Mouse Brain	ABA	Standard	None	Brain	Mus musculus	None	None
	Access to Biological Collection Data	ABCD	Standard	Biodiversity Biology Life Sciences	None	All	GBIF Atlas of Living Australia IPT - GBIF Australia Repository GBIF Spain IPT - GBIF Spain Repository Canadensys	ABCD EF ABCDDN
	Access to Biological Collection Databases Extended for Geosciences	ABCD EFG	Standard	Geology Paleontology Soil Sciences	None			

Standards

1294 Standards

Terminology Artifact	728
Model/Format	383
Reporting Guideline	159
Identifier Schema	10
Metrics	14

FAIRSharing; FAIRsharing as a community approach to standards, repositories and policies, April 2019

Reusable

R	Reusable
R1	meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1	(meta)data are released with a clear and accessible data usage license
R1.2	(meta)data are associated with detailed provenance
R1.3	(meta)data meet domain-relevant community standards

Reusable

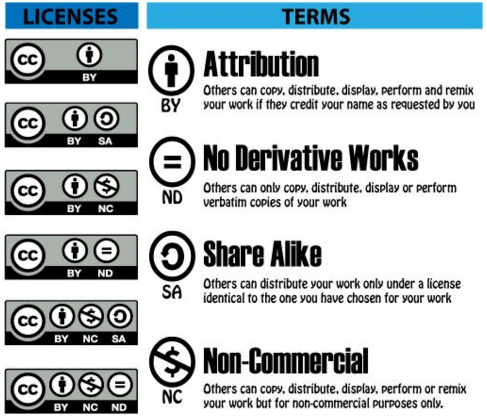
R	Reusable
R1	meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1	**(meta)data are released with a clear and accessible data usage license**
R1.2	(meta)data are associated with detailed provenance
R1.3	(meta)data meet domain-relevant community standards

Data usage license should be clear!

Reusers need clear signals from researchers on what they can and cannot do with their research data.

Best practice

1. Check for funder/data repository/local policy/institution obligations
2. If your obligations are non-exclusive, consider multiple-licensing (different versions)
2. Avoid bespoke licenses



Others can copy, distribute, display, perform and remix your work if they credit your name as requested by you

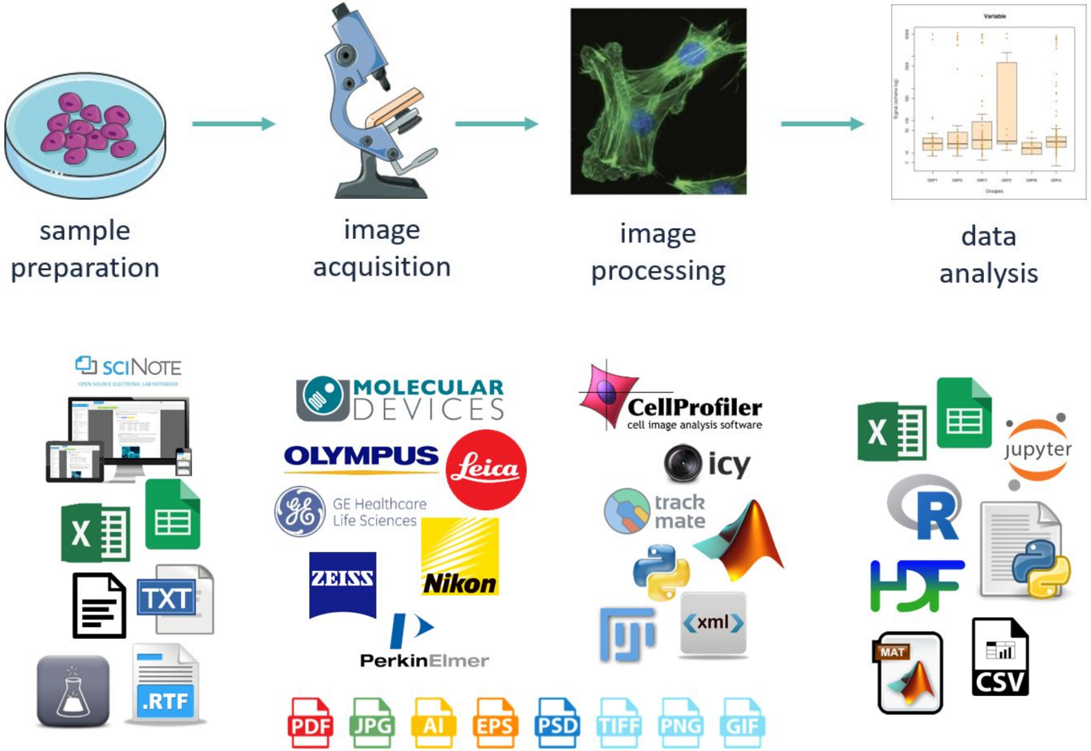
Others can only copy, distribute, display or perform verbatim copies of your work

Others can distribute your work only under a license identical to the one you have chosen for your work

Others can copy, distribute, display, perform or remix your work but for non-commercial purposes only.

How to license Research Data

My story: towards FAIR cell migration data



sample preparation

image acquisition

image processing

data analysis

[Taking Aim at Moving Targets in Computational Cell Migration, Masuzzo *et al.*, 2016]

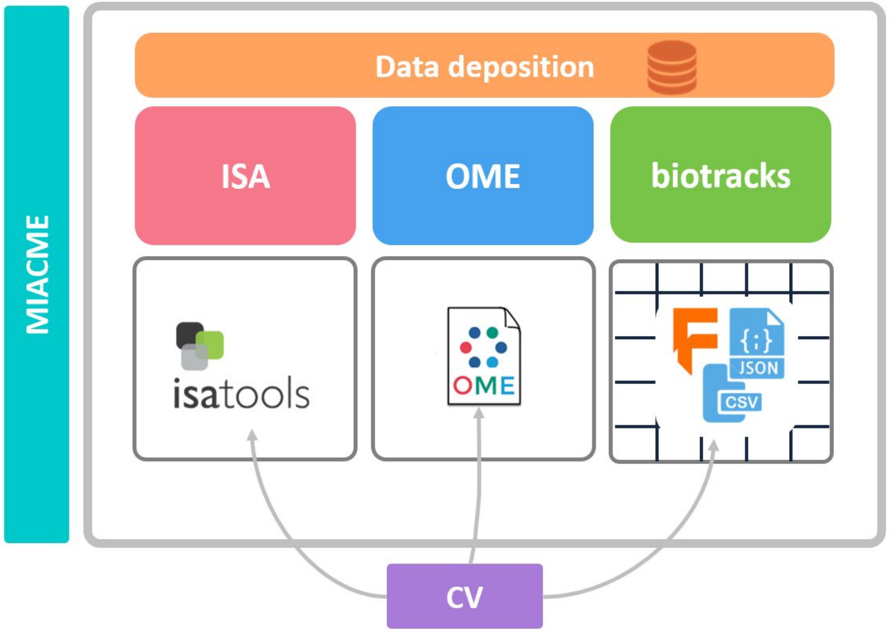
My story: towards FAIR cell migration data

MIACME: Minimum Information About a Cell Migration Experiment

ISA: Investigation Study Assay

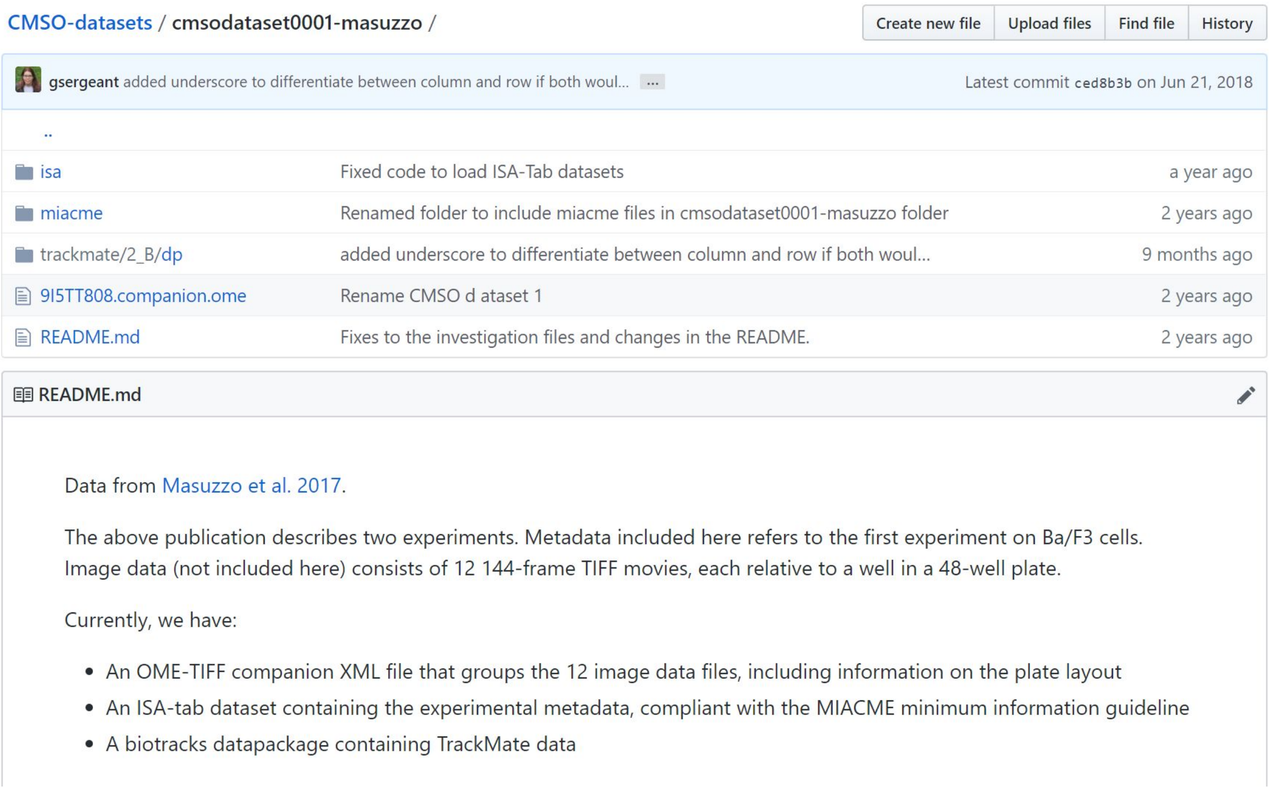
OME: Open Microscopy Environment

biotracks: specialises the Tabular Data Package container format



Gonzales, Masuzzo et al., 2019; Frictionless Data

A FAIR cell migration dataset



CMSO-datasets / cmsodataset0001-masuzzo /

gsergeant added underscore to differentiate between column and row if both woul... … Latest commit ced8b3b on Jun 21, 2018

..		
isa	Fixed code to load ISA-Tab datasets	a year ago
miacme	Renamed folder to include miacme files in cmsodataset0001-masuzzo folder	2 years ago
trackmate/2_B/dp	added underscore to differentiate between column and row if both woul...	9 months ago
9I5TT808.companion.ome	Rename CMSO d ataset 1	2 years ago
README.md	Fixes to the investigation files and changes in the README.	2 years ago

README.md

Data from Masuzzo et al. 2017.

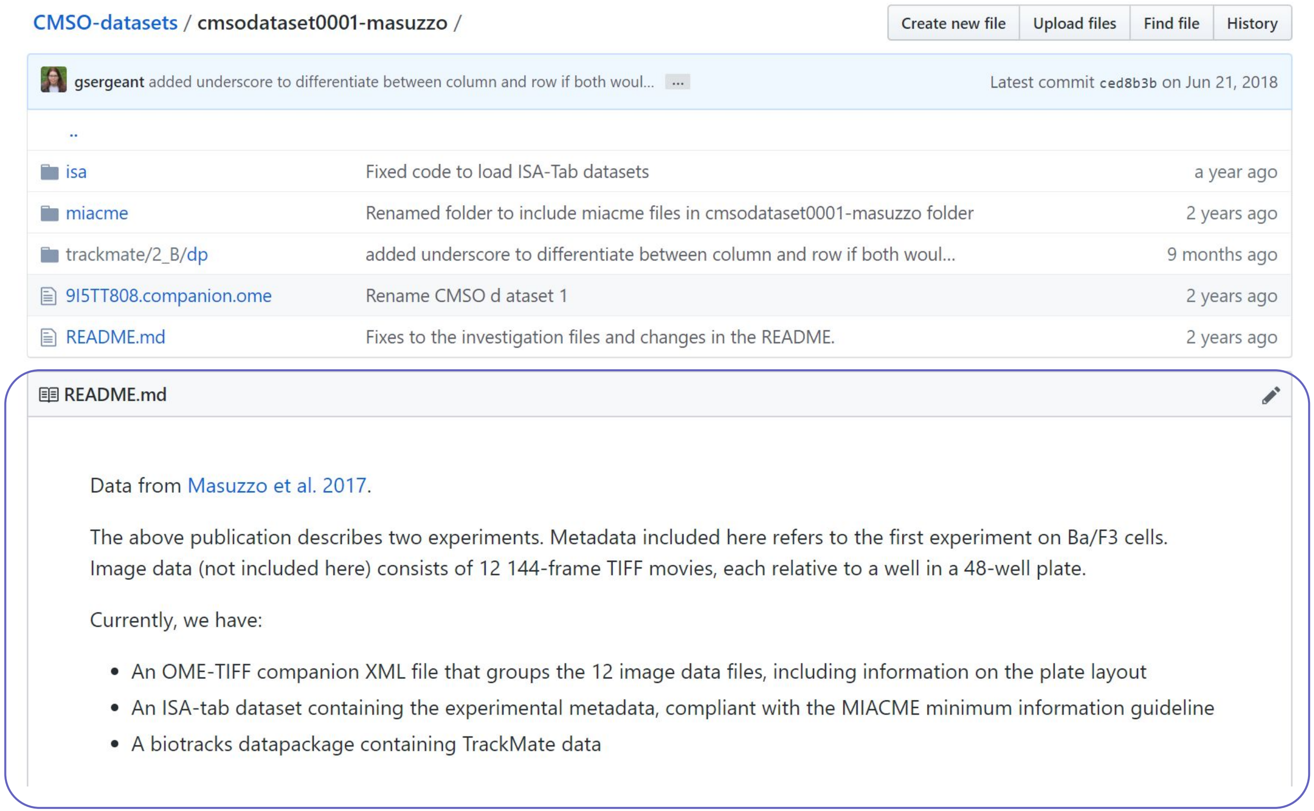
The above publication describes two experiments. Metadata included here refers to the first experiment on Ba/F3 cells. Image data (not included here) consists of 12 144-frame TIFF movies, each relative to a well in a 48-well plate.

Currently, we have:

- An OME-TIFF companion XML file that groups the 12 image data files, including information on the plate layout
- An ISA-tab dataset containing the experimental metadata, compliant with the MIACME minimum information guideline
- A biotracks datapackage containing TrackMate data

CMSO dataset

A FAIR cell migration dataset



CMSO-datasets / cmsodataset0001-masuzzo /

gsergeant added underscore to differentiate between column and row if both woul... ... Latest commit ced8b3b on Jun 21, 2018

..		
isa	Fixed code to load ISA-Tab datasets	a year ago
miacme	Renamed folder to include miacme files in cmsodataset0001-masuzzo folder	2 years ago
trackmate/2_B/dp	added underscore to differentiate between column and row if both woul...	9 months ago
9I5TT808.companion.ome	Rename CMSO d ataset 1	2 years ago
README.md	Fixes to the investigation files and changes in the README.	2 years ago

README.md

Data from Masuzzo et al. 2017.

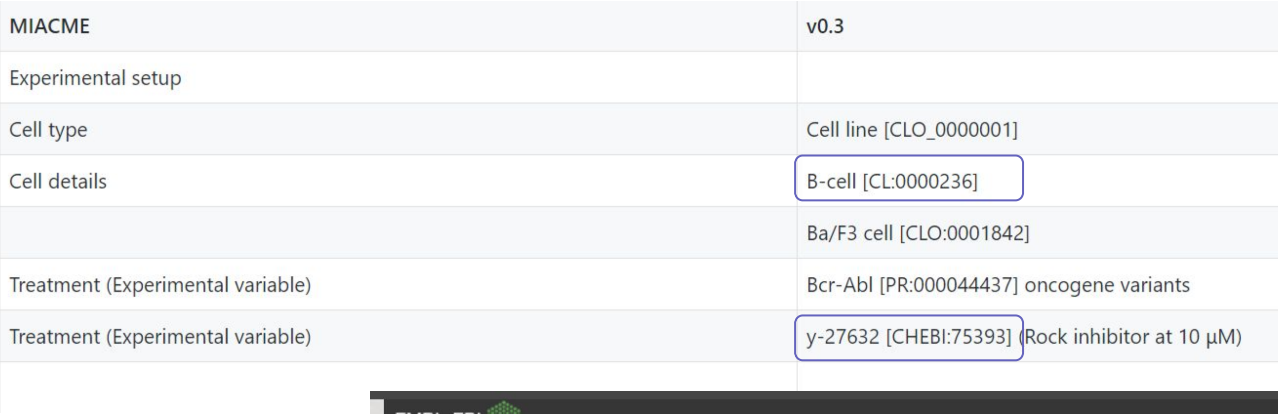
The above publication describes two experiments. Metadata included here refers to the first experiment on Ba/F3 cells. Image data (not included here) consists of 12 144-frame TIFF movies, each relative to a well in a 48-well plate.

Currently, we have:

- An OME-TIFF companion XML file that groups the 12 image data files, including information on the plate layout
- An ISA-tab dataset containing the experimental metadata, compliant with the MIACME minimum information guideline
- A biotracks datapackage containing TrackMate data

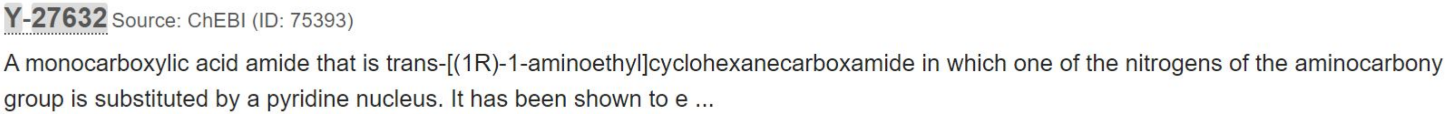
Annotate, annotate, annotate!

A FAIR cell migration dataset



MIACME	v0.3
Experimental setup	
Cell type	Cell line [CLO_0000001]
Cell details	B-cell [CL:0000236]
	Ba/F3 cell [CLO:0001842]
Treatment (Experimental variable)	Bcr-Abl [PR:000044437] oncogene variants
Treatment (Experimental variable)	y-27632 [CHEBI:75393] (Rock inhibitor at 10 µM)

Use standard ontologies and controlled vocabularies!



Y-27632 Source: ChEBI (ID: 75393)

A monocarboxylic acid amide that is trans-[(1R)-1-aminoethyl]cyclohexanecarboxamide in which one of the nitrogens of the aminocarbony group is substituted by a pyridine nucleus. It has been shown to e ...

A FAIR cell migration dataset

```
<Image ID="Image:2" Name="9I5TT808_F00000012.tif">
   <AcquisitionDate>2013-03-25T22:05:53</AcquisitionDate>
   <Description>Camera 1376x1038, Controler present, Lightsource
off</Description>
   <Pixels BigEndian="false" DimensionOrder="XYCZT" ID="Pixels:0"
Interleaved="false" PhysicalSizeX="0.322000000022352"
PhysicalSizeXUnit="µm" PhysicalSizeY="0.322000000022352"
PhysicalSizeYUnit="µm"
SignificantBits="16"
SizeC="1" SizeT="144" SizeX="1376" SizeY="1038" SizeZ="1"
Type="uint16">
     <Channel ID="Channel:0:0" SamplesPerPixel="1">
       <LightPath/>
     </Channel>
     <TiffData FirstC="0" FirstT="0" FirstZ="0" IFD="0" PlaneCount="1">
       <UUID
FileName="9I5TT808_F00000012.tif">urn:uuid:90e9dc02-94db-4e32-
9759-2780958e6c6d</UUID>
     </TiffData>
```

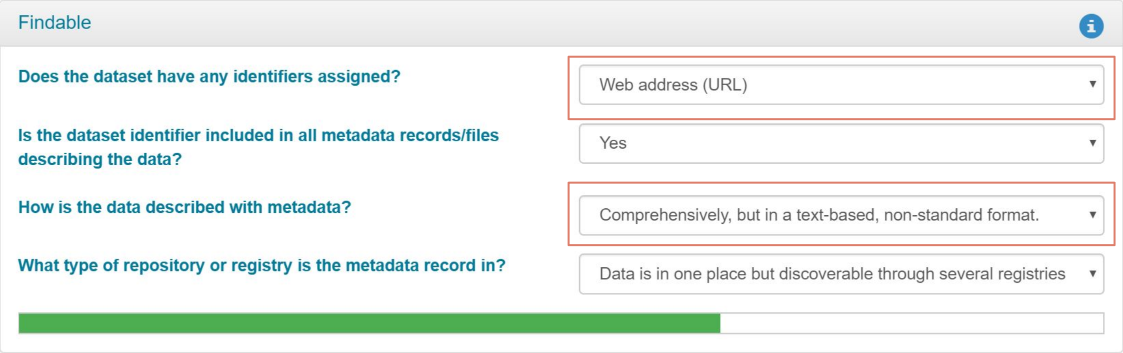
Expose as rich metadata as possible!

Note

Some data elements, for instance images and 'raw data' can not always be made machine-processable.

Being published with FAIR metadata is of very high value in its own right.

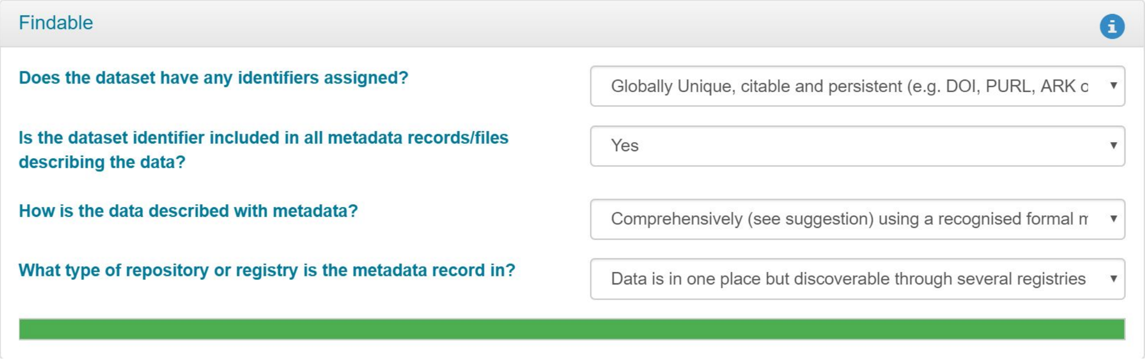
Different levels of FAIR-ness can exist



Findable

Does the dataset have any identifiers assigned?	Web address (URL)
Is the dataset identifier included in all metadata records/files describing the data?	Yes
How is the data described with metadata?	Comprehensively, but in a text-based, non-standard format.
What type of repository or registry is the metadata record in?	Data is in one place but discoverable through several registries

A little step can make a big difference



Findable

Does the dataset have any identifiers assigned?	Globally Unique, citable and persistent (e.g. DOI, PURL, ARK c
Is the dataset identifier included in all metadata records/files describing the data?	Yes
How is the data described with metadata?	Comprehensively (see suggestion) using a recognised formal n
What type of repository or registry is the metadata record in?	Data is in one place but discoverable through several registries

https://www.ands-nectar-rds.org.au/fair-tool

Ethical considerations and legal obligations guide the way

Data Management or Data Stewardship ?

With appropriate **data management planning** much sensitive and proprietary data can be shared, reused, and FAIR.

The metadata can almost always be shared.

Guidance and best practices for sharing sensitive data are necessarily region-specific because of local regulations.
As general guidance, the EU RESPECT code for Ethics and data protection highlights 3 key aspects:

1) **Upholding scientific standards**
when formulating your research questions, you do not pre-determine or prejudice the outcome through your choice of questions or actions

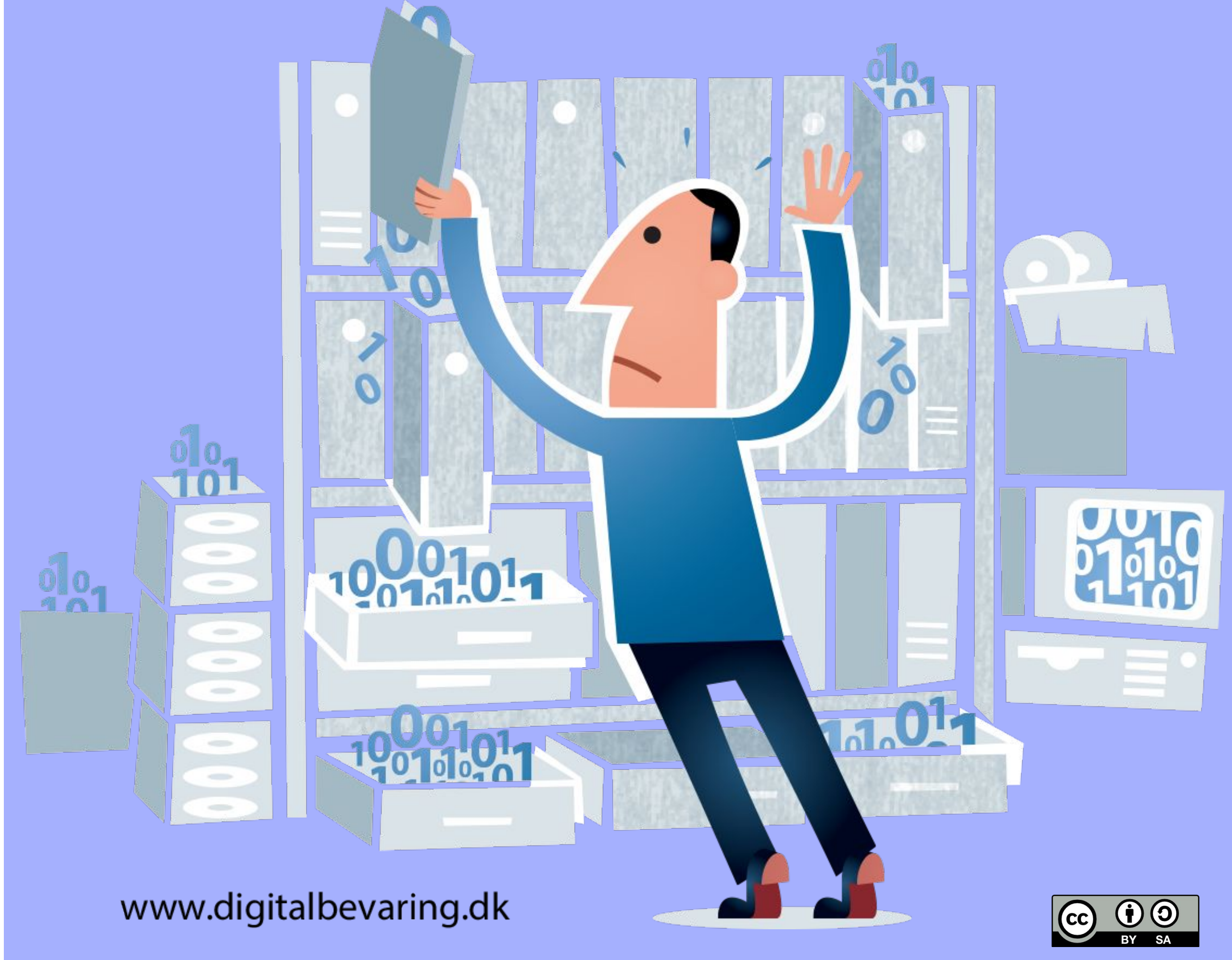
2) **Compliance with the law**
be aware of all the relevant national and international laws that may affect your research project. With collaborative projects which cross legal borders, this may involve various laws. Ones of particular relevance will be in regards to data protection and intellectual property.

3) **Avoidance of social and personal harm**
your research project should be designed responsibly and should consider participants throughout. For example, participation in the research project should be voluntary and on the basis of fully informed consent

Ethics and Data Protection

A goal without a plan is just a wish

[Antoine de Saint-Exupéry (1900 -1944)]



What is a Data Management Plan (DMP)?

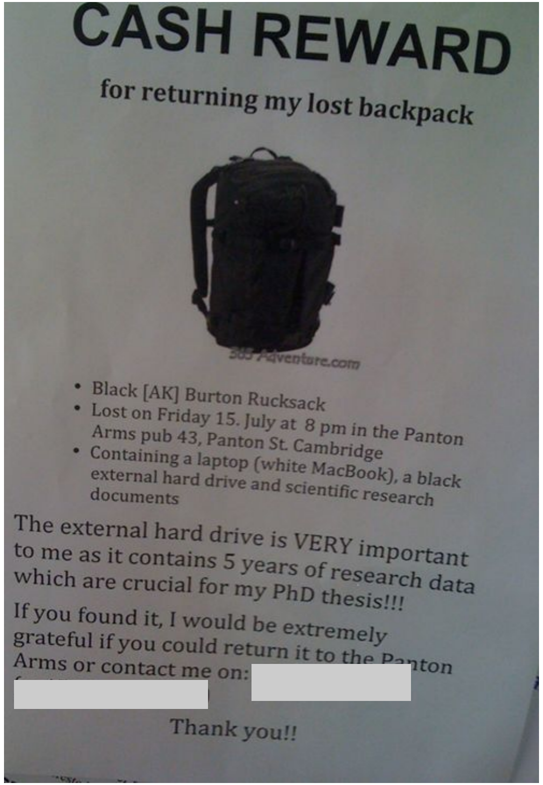
A DMP is a term describing how you organize, structure, store and care for the information used in a research project

How do you look at the data on a day-by-day base?

Organize your data, store them and back them up.
Choose the right file formats.
Document your data.

What happens to the data after the research project finishes?

What data do you need to keep (and share)?
What data must not be kept (and shared)?
How are you going to achieve long-term data storage/access?



CASH REWARD

for returning my lost backpack

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

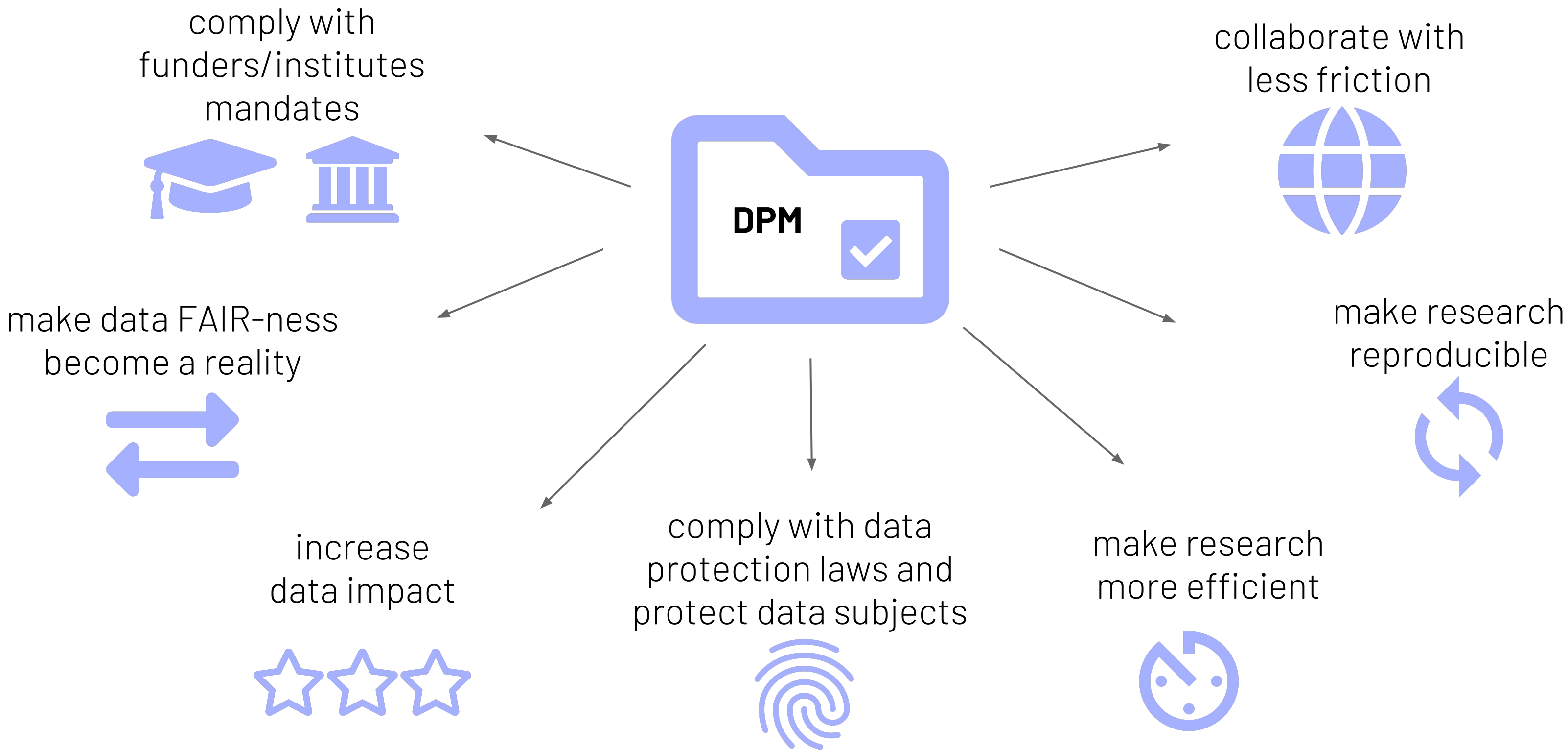
The external hard drive is VERY important to me as it contains 5 years of research data which are crucial for my PhD thesis!!!

If you found it, I would be extremely grateful if you could return it to the Panton Arms or contact me on: [redacted] [redacted]

Thank you!!

Murray-Rust, Peter - Why YOU need a data management plan

Why should you develop a DMP?



comply with funders/institutes mandates

collaborate with less friction

make data FAIR-ness become a reality

make research reproducible

increase data impact

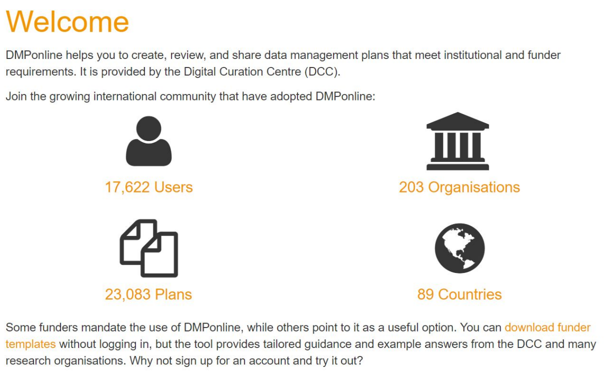
comply with data protection laws and protect data subjects

make research more efficient

How do you create a DMP?

My personal message is: don't be afraid, go to your desk and try!

DPMOnline developed by the DCC has free-to-download public templates + your university might have a local version of the tool, which means you can use your institutional account to sign in and the templates available will make sure you don't miss out on relevant info/fields.



Welcome

DMPonline helps you to create, review, and share data management plans that meet institutional and funder requirements. It is provided by the Digital Curation Centre (DCC).

Join the growing international community that have adopted DMPonline:

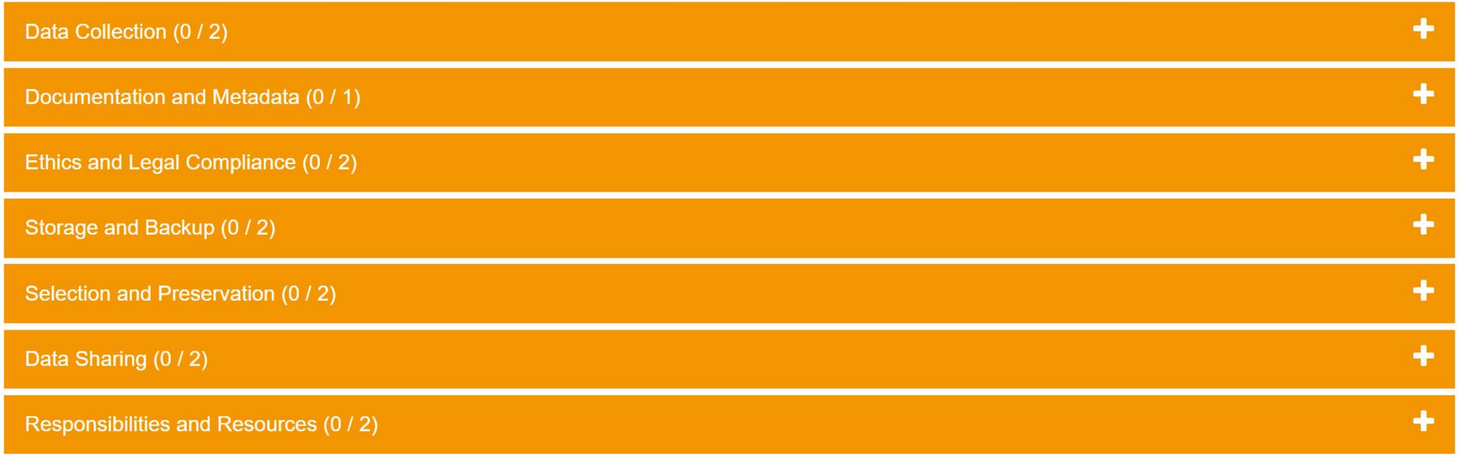
17,622 Users

203 Organisations

23,083 Plans

89 Countries

Some funders mandate the use of DMPonline, while others point to it as a useful option. You can download funder templates without logging in, but the tool provides tailored guidance and example answers from the DCC and many research organisations. Why not sign up for an account and try it out?



Data Collection (0 / 2) +

Documentation and Metadata (0 / 1) +

Ethics and Legal Compliance (0 / 2) +

Storage and Backup (0 / 2) +

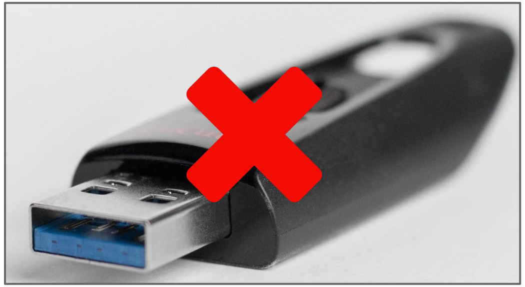
Selection and Preservation (0 / 2) +

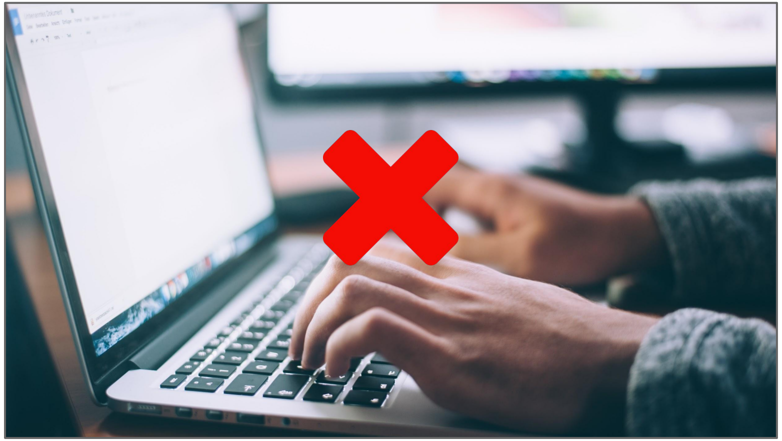
Data Sharing (0 / 2) +

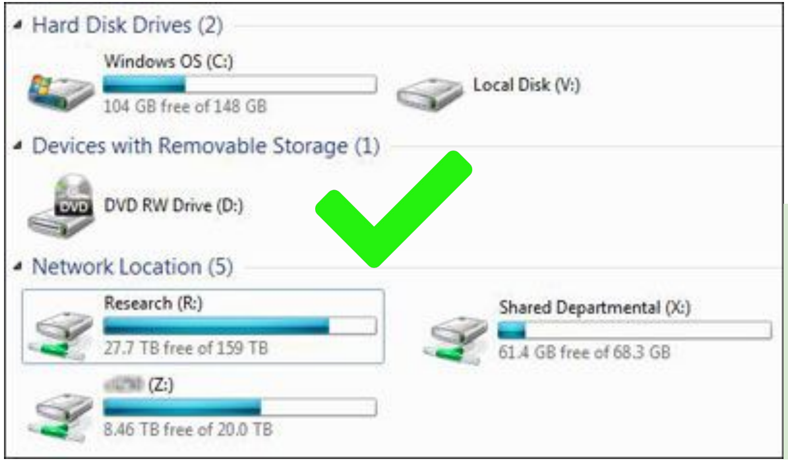
Responsibilities and Resources (0 / 2) +

Data Management Checklist, https://dmponline.dcc.ac.uk

Where are you going to store your data?







Hard Disk Drives (2)

Windows OS (C:)
104 GB free of 148 GB

Local Disk (V:)

Devices with Removable Storage (1)

DVD RW Drive (D:)

Network Location (5)

Research (R:)
27.7 TB free of 159 TB

Shared Departmental (X:)
61.4 GB free of 68.3 GB

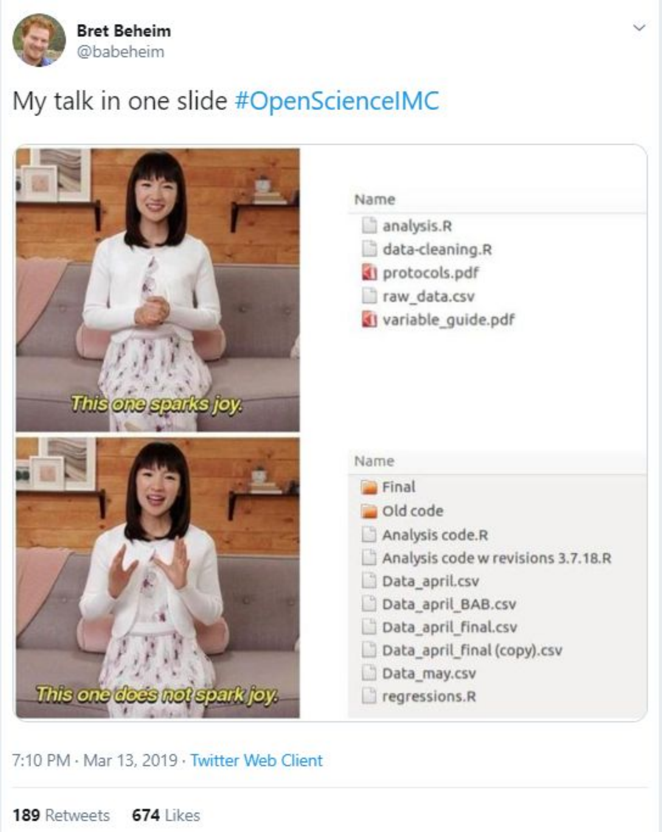
[redacted] (Z:)
8.46 TB free of 20.0 TB

central file shares managed by the ICT Department of your University

live snapshots, backups!
prevent disasters!

[11 ways to avert a data-storage disaster]()

How are you going to organize your data?



Bret Beheim
@babeheim

My talk in one slide #OpenScienceIMC

This one sparks joy.

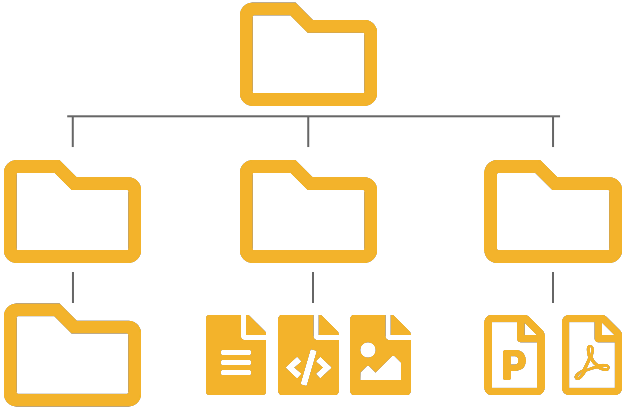
Name
- analysis.R
- data-cleaning.R
- protocols.pdf
- raw_data.csv
- variable_guide.pdf

This one does not spark joy.

Name
- Final
- Old code
- Analysis code.R
- Analysis code w revisions 3.7.18.R
- Data_april.csv
- Data_april_BAB.csv
- Data_april_final.csv
- Data_april_final (copy).csv
- Data_may.csv
- regressions.R

7:10 PM · Mar 13, 2019 · Twitter Web Client

189 Retweets 674 Likes



use directories and folders hierarchy

name files with a bit of common sense: no special characters, no capitals, no spaces

know what your file is before you double click it!

What formats and what software will you use?

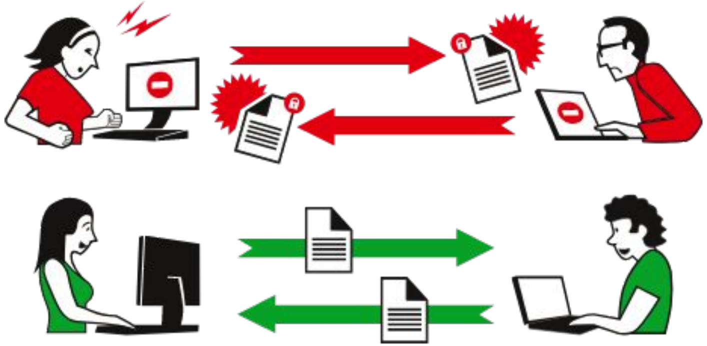
During your research you will most likely use formats that fit your scope, workflow and methodology

After research is completed, you should look for formats that are easy to share and re-use and that may last longer into the future

- open specifications
- widely used formats
- uncompressed
- ASCII formats
- exchange formats

tip:check if software you are using has an option to export into a more suitable format for sharing or long-term reuse

(sometimes 'crazy' formats are simply ASCII-based, and can be saved and then opened as regular .txt files!)



https://en.wikipedia.org/wiki/List_of_open_formats, 1st image CC-BY-SA Regards Citoyens, 2nd image ThinkStock

THANKS!

You can find me at:

@pcmasuzzo

paola.masuzzo@gmail.com