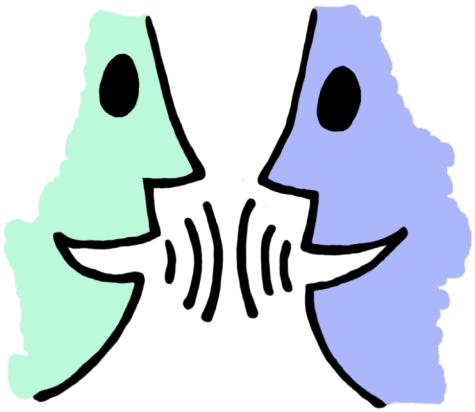
Data sharing and Backup

Anne Pajon 26th February 2019
Mark Fernandes (Updated Nov 2019)

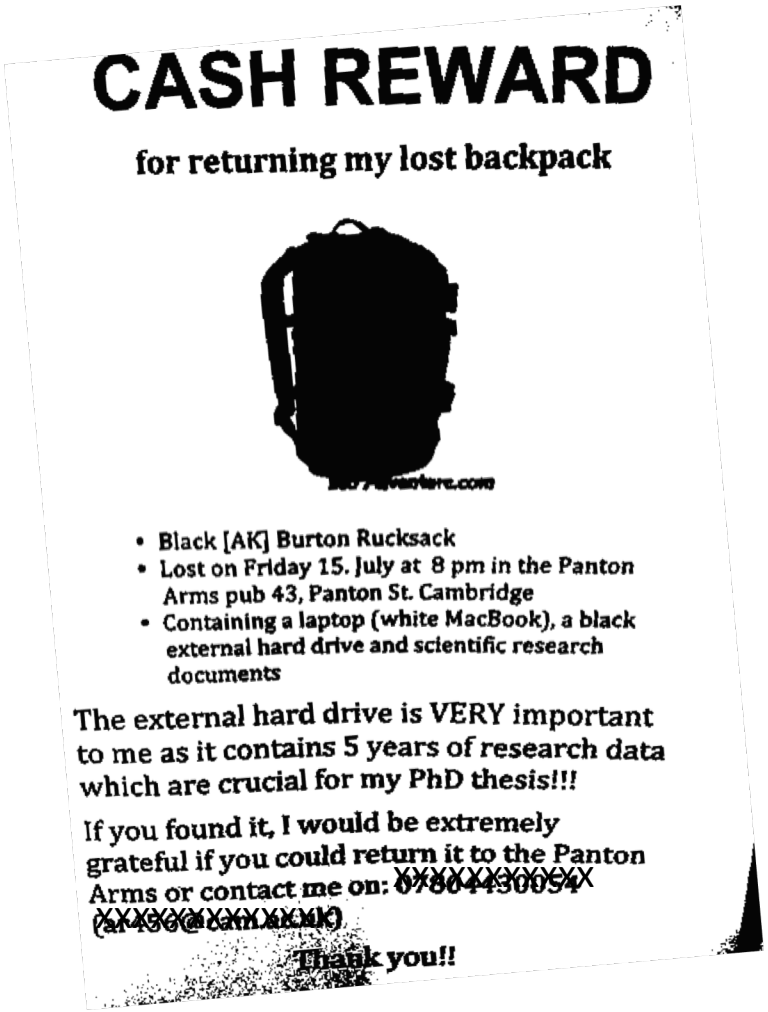
How do you **manage** your research data?

- What kind of research data do you have?
- Do you do any data **backup**?
- How **often**?
- How do you **share** files/data with collaborators?



To avoid data ***disasters*** ...

What would you do if you'd lose your data tomorrow?



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: 07864430054 XXXXXXXXXXX

(ar456@cam.ac.uk) XXXXXXXXXXXXXX

Thank you!!

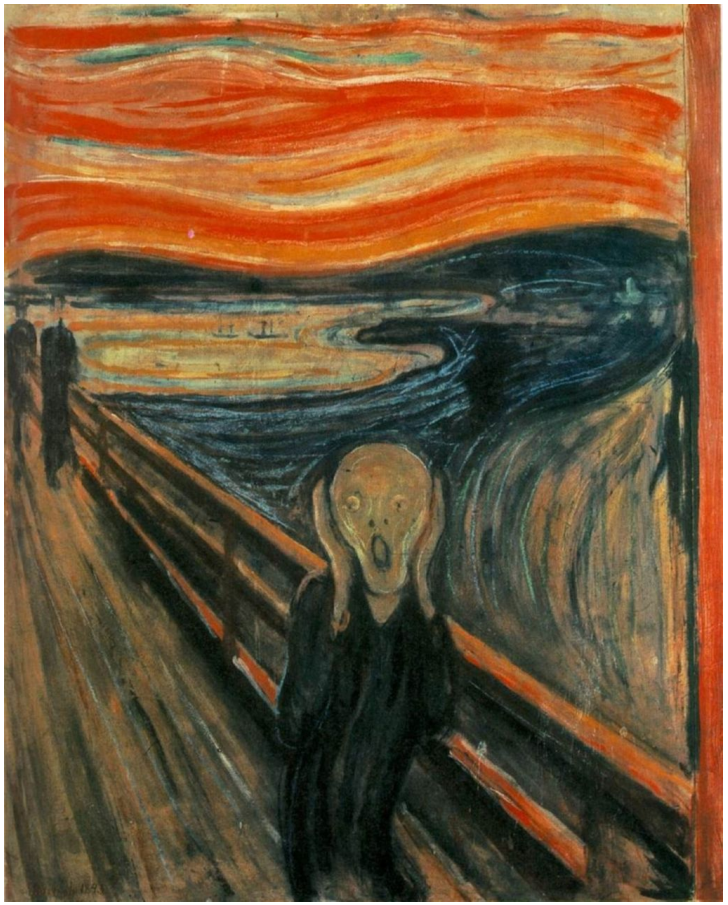
https://blogs.ch.cam.ac.uk/pmr/2011/08/01/why-you-need-a-data-management-plan/

What if?

Cancer Research UK – University of Manchester – 27 April 2017

What would you do if you'd lose your data tomorrow?

- Your laptop got stolen
- Your office/house burnt
- Your USB stick is lost
- Your portable hard disk is damaged
- Your data in Dropbox disappeared



https://en.wikipedia.org/wiki/The_Scream

As stated in the first talk:

Never work directly on the raw data

Leave it intact

Always **make a copy**, and work on the copy

Data backup

*At least **2** backups at **2** different locations*

External disks

Online backup

Dropbox Google Drive OneDrive

Servers

Department
College
IT

	External disks	Online backup	Servers
🙂	Cheap £10-15 / TB (1024GB)	Accessibility Free (limit)	Managed by experts
☹️	Failure rate 1.5%/year	Personal data Hacking	Moving between institutions

Data backup

Manual

Copying files to relevant folders

Dropbox Google Drive

OneDrive

Copying files to relevant folders

Automated

- Install software
e.g. Time machine (Mac users)
Syncback(Windows)
https://www.2brightsparks.com

- RAID technology

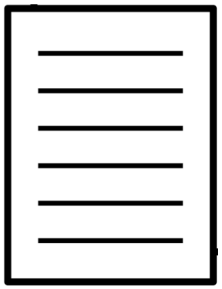
- Checksums

Automatically upload files to the cloud when any changes are saved

If manual … how often?

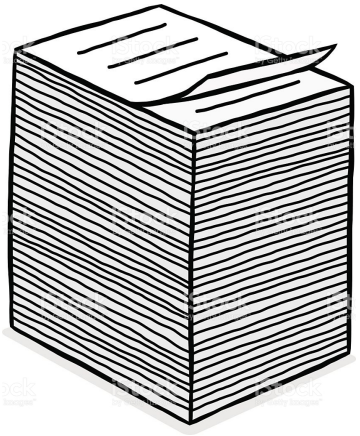
How much would you be willing to lose?

1 day



1 week

1 month-year



*Software allows you to set up **backup time** automatically*

Data backup and file sharing

	Dropbox	Google Drive	OneDrive
Space/price	2 GB (free) 2 TB (£96/year)	15 GB (free) 2 TB (£80/year)	5 GB (free) 1 TB (£60)
File history and recovery	Yes, unlimited	Yes	Last 90 days
File size limit	None	Up to 5 TB	15 GB
Support	UIS	UIS	UIS
OS	Windows, Mac, Linux, Android, iOS	Windows, Mac, Android, iOS	Windows, Mac, Android, iOS
Accessibility	Sync anywhere on any devices	Live editing	Integration with Microsoft Office

Beware of the cloud…

reliability

Is the cloud storage supplier reliable?

security

Have you lost control over your data?

cost

Is there hidden cost?

Why data sharing is important?

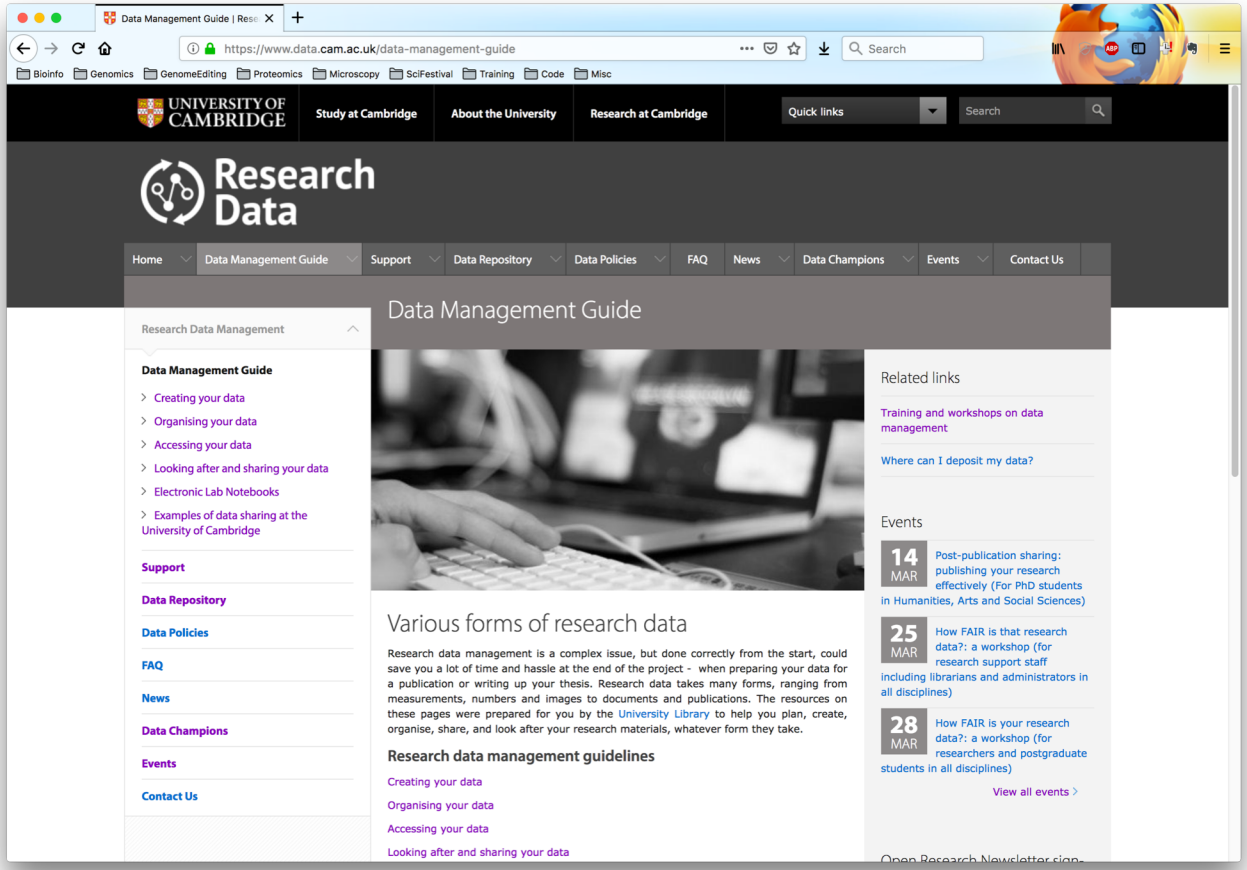
Data should be shared to move our knowledge forward.

Research Data Policies

http://www.data.cam.ac.uk/data-management-guide

Most research funders have also introduced **policies** on research data management.

The general expectation is that publicly funded research data are a **public good**, and should be made **openly available** with as few restrictions as possible.



What data to share?

- Data & metadata
  - Raw
  - Processed
- Code (software and scripts)
- Methods
- Papers
  - results and figures

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

https://doaj.org/

- What about non-positive results?
  - RIO - Research Ideas and Outcomes

https://riojournal.com/

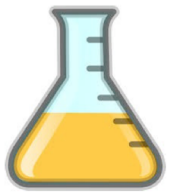
  - Would you consider Pre-Print for your draft manuscripts?

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

https://www.biorxiv.org/

When to share?



Close data

Open data

Initial experiments

Method optimisation

Answering biological questions

Write up story

Ideas

Paper

Tests

First scripts

Robust analysis pipeline

Figures Tables

Document computational method

Private code

Public code

Gather information early

- **Think** of data submission from the start
- Collect **metadata** before submission
- Keep a **ReadMe** file about your project
- Consider using an **E**lectronic **L**ab **N**otebook

How to share your data?

Store, describe and deposit your data in suitable and trusted public data repositories and add a link to your data in your publication.

Repositories for data

- Discipline specific
  - Registry of Research Data Repositories http://www.re3data.org/
  - EMBL-EBI services https://www.ebi.ac.uk/services
- General purpose
  - Zenodo https://zenodo.org/
  - FigShare https://figshare.com

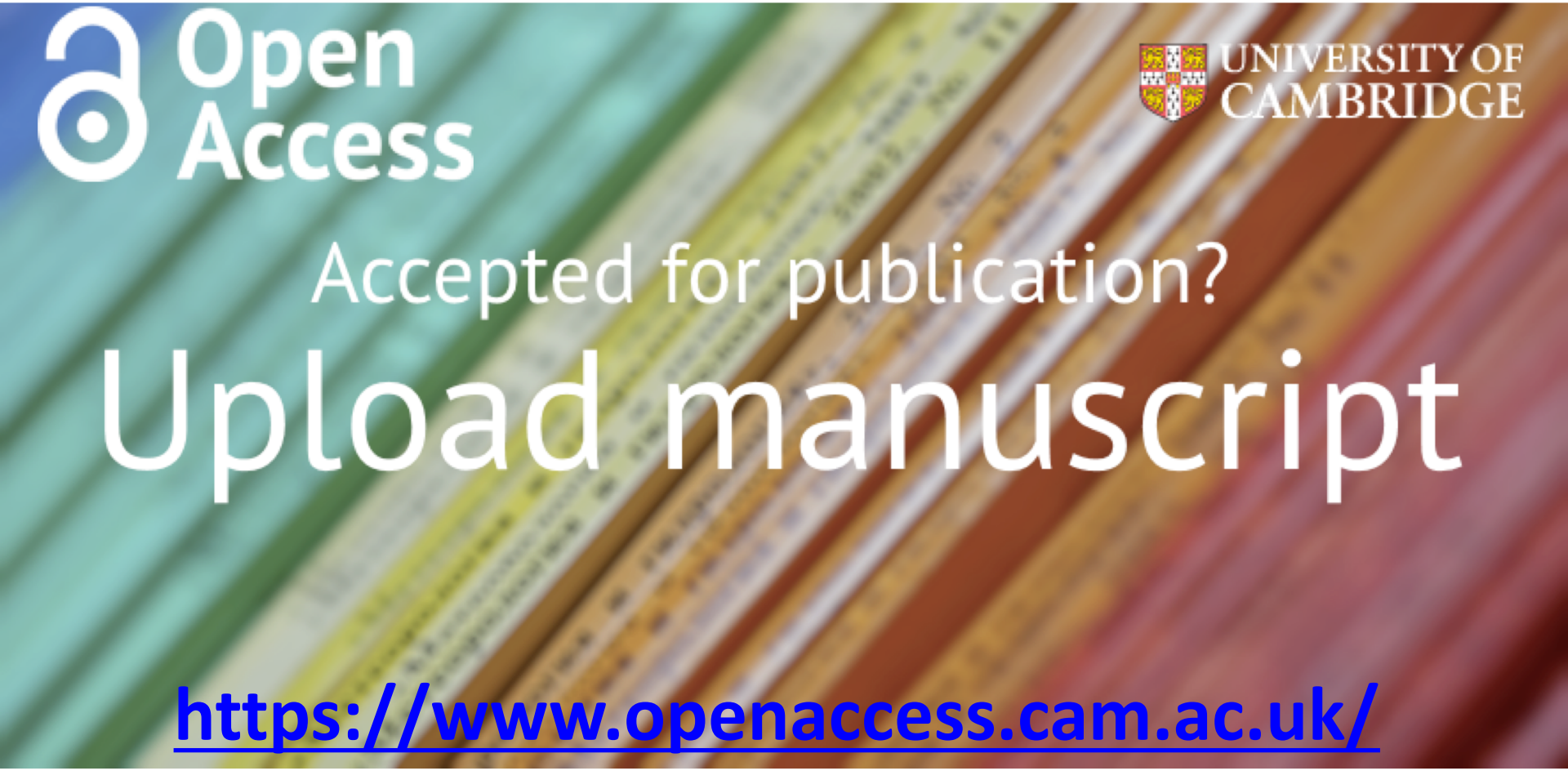
Repositories for code

- GitHub https://github.com
- GitLab https://gitlab.com
- Bitbucket https://bitbucket.org

Zenodo assigns a Digital Object Identifier (DOI) to make the upload easily and uniquely citable, with GitHub integration to enable tracking of each release.

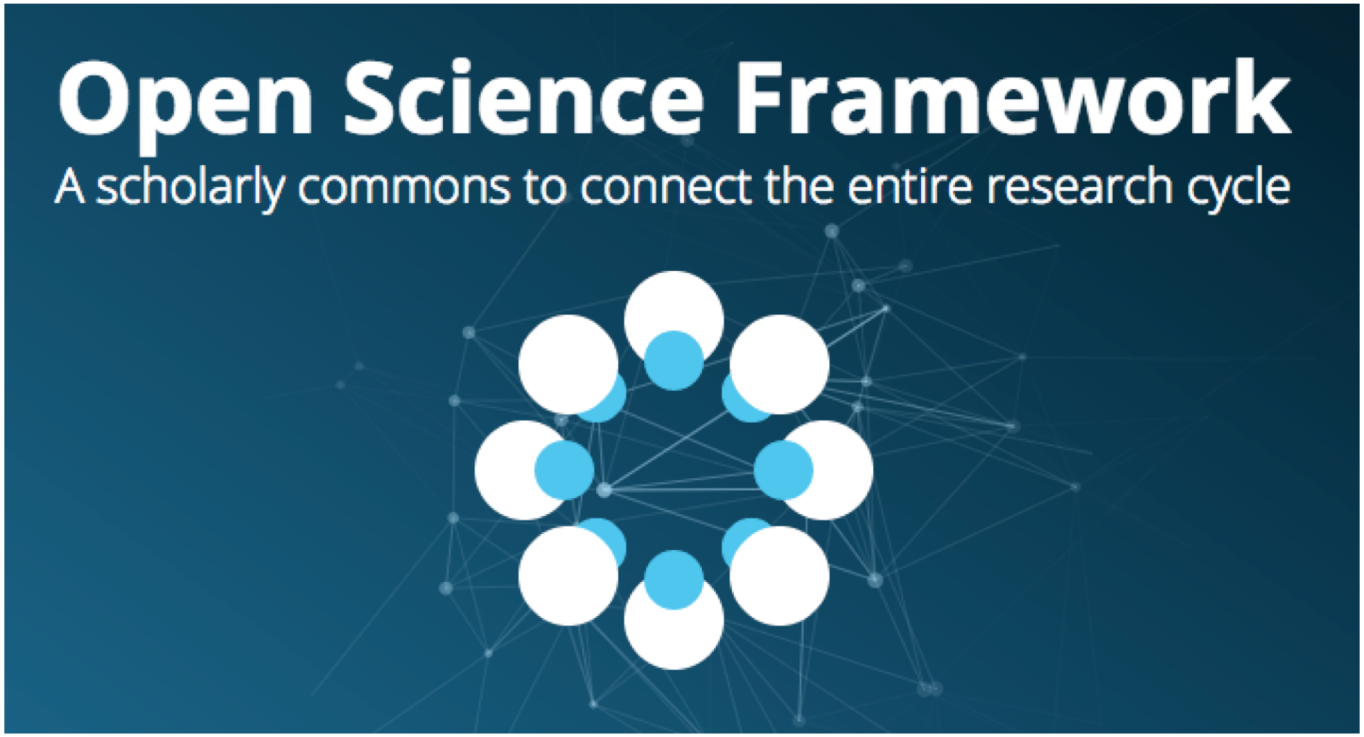
How to make your publication Open Access?

The Open Access Team will check your funder and journal policies and advise on how to comply with Open Access requirements.



Open Science Framework

Cloud-based management for your projects @ https://osf.io/

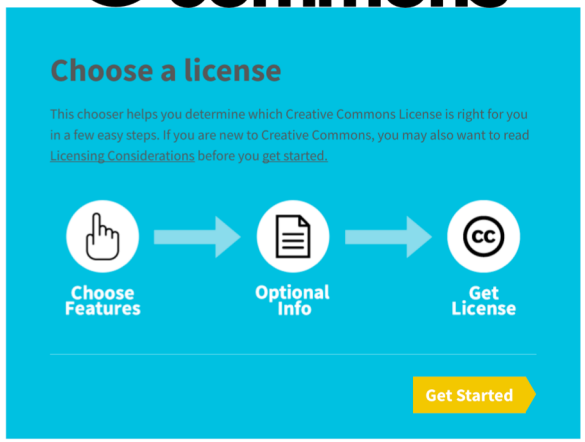


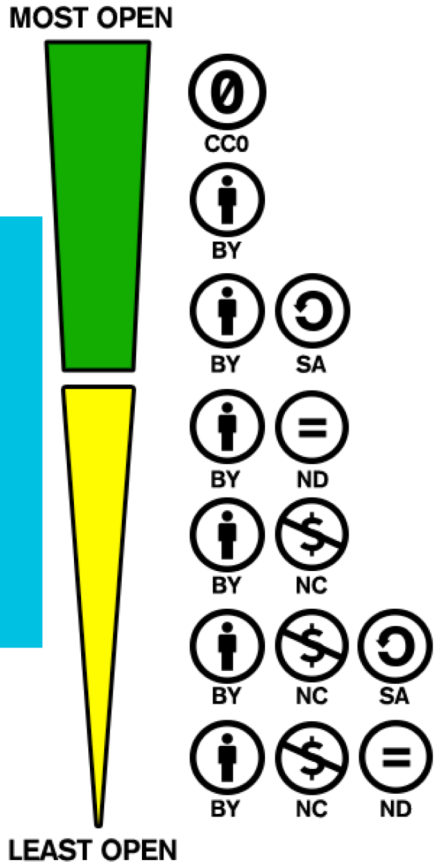
Under which license?

Creative work is under exclusive copyright by default.

Share your data with one of the Creative Commons licenses

https://creativecommons.org/

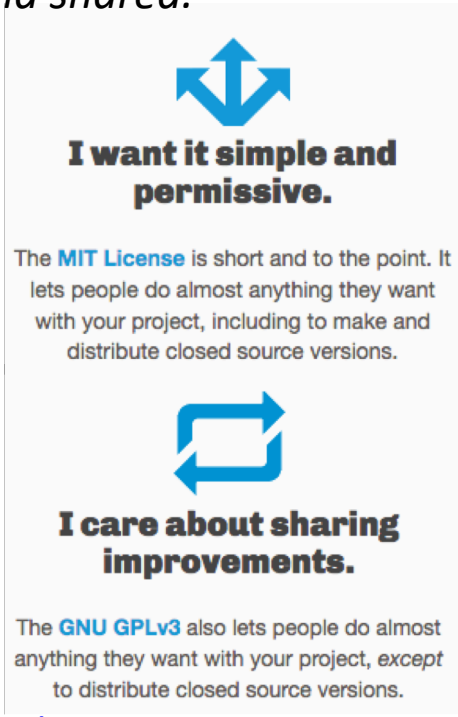




Share your code with one of the open source licenses

https://opensource.org/licenses

They allow software to be freely used, modified, and shared.



https://choosealicense.com

Open Access Policy

As a CRUK-funded researcher, if you have an original primary article accepted for publication in a peer-reviewed journal, **you must ensure** that:

- An electronic copy of the final, published form of **your paper is available on Europe PubMed Central** (Europe PMC) as soon as possible and no later than 6 months after publication.

- If you've paid an article processing charge (APC) for the Europe PMC deposit, your paper must be published with **a CC-BY license**, so that it may be freely copied and reused, providing that the original authors are properly credited. *Other licenses will not be compliant with your Grant Conditions.*

- The journal you publish in must be published by a **publisher who has agreed to the COAF/Wellcome Trust publisher requirements**.

https://www.cancerresearchuk.org/sites/default/files/policy_on_open_access.pdf

Conclusion

- Always **make a copy** of your data
- **Backup** your data at least **twice** at two different locations
- Document your process using a **ReadMe** file
- Ideally most **data should be shared**
  - Sharing is essential for all publicly funded research
  - Share as early as possible
  - Using suitable repositories and DOI
- Share your work under a Creative Commons

or Open Source **license**

Quiz

https://frama.link/manage-data-quiz

Managing your Research data Quiz

*Required

Would you work directly on raw data or make a copy of it? *

- ◯ Work directly on raw data
- ◯ Make a copy and work on it
- ◯ Other: _____

Which file should I create to document my data and processes? *

- ◯ License file
- ◯ ReadMe file
- ◯ ToDo file
- ◯ Other: _____

Which inconsistencies have you encounter in data? *

- ☐ Different values for similar information (e.g. female/male)