

We will be using an electronic whiteboard (Etherpad).

Please log onto your computer and access:

<https://public.etherpad-mozilla.org/p/2019-2-26-cruk-ci-myrd>

Data sharing and Backup

Anne Pajon

26th February 2019

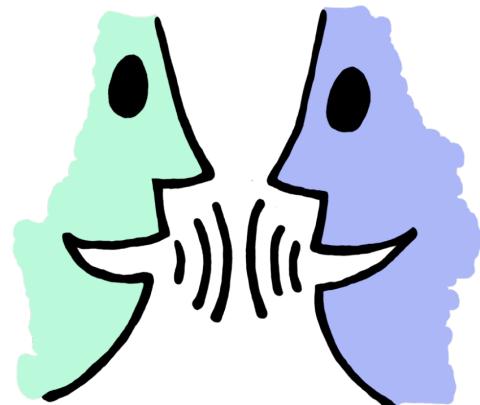


This work is licensed under a [Creative Commons Attribution 4.0 International License](#).

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

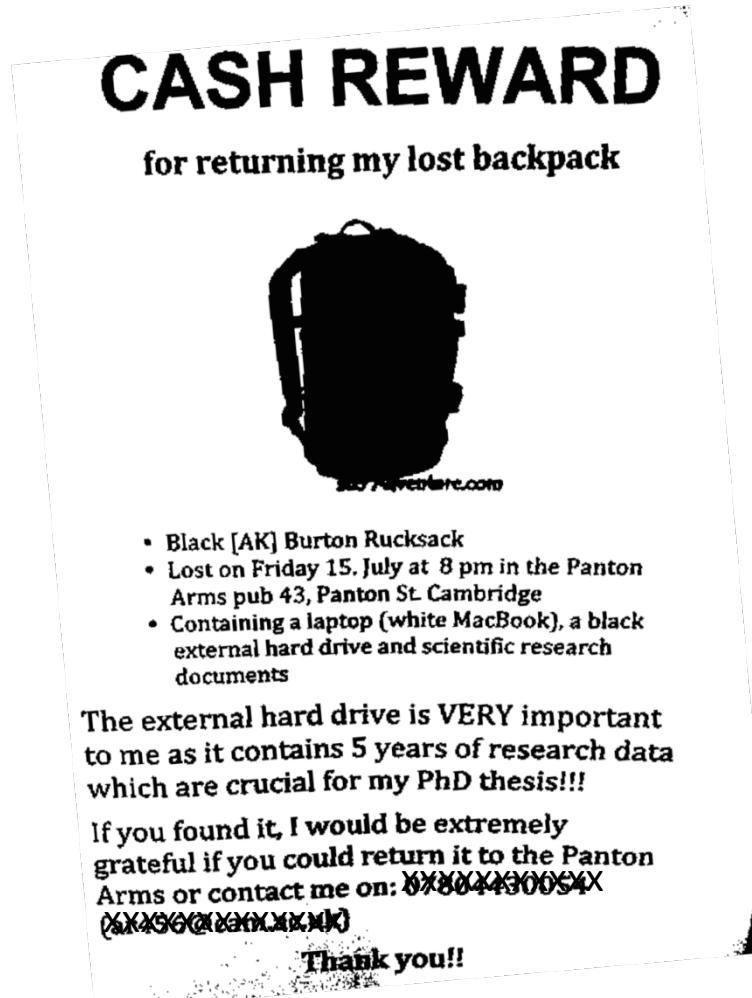
Talk to the person next to you for 5 min and exchange about:

- 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?



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

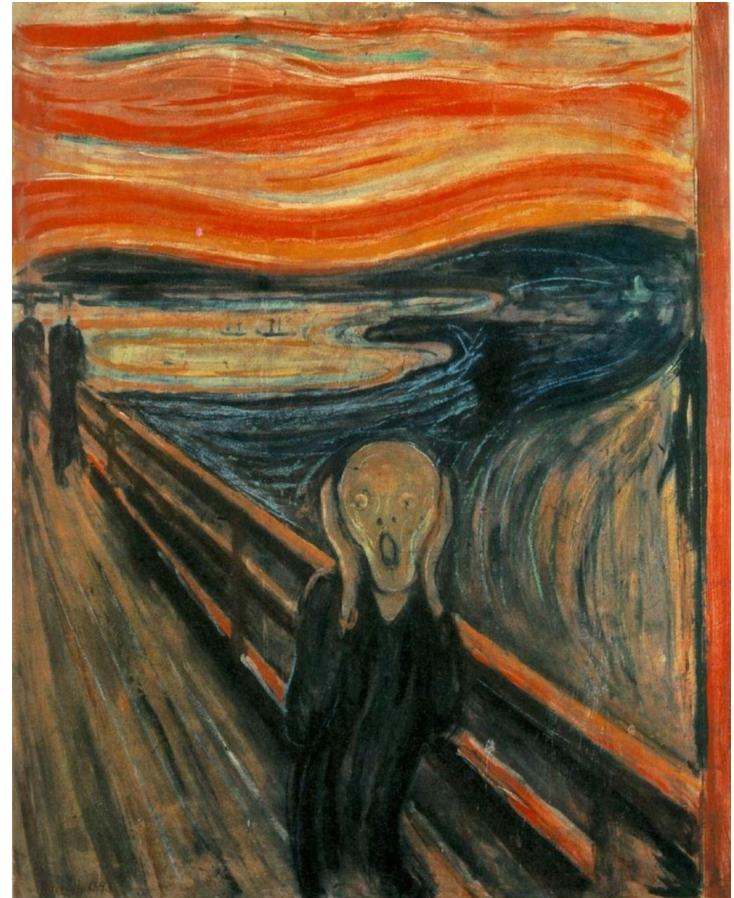
What if?



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

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



Servers

Department
College
IT



Cheap
£10-15 / TB (1024GB)



Failure rate
1.5%/year

Accessibility
Free (limit)

Personal data
Hacking

Managed by
experts

Moving between
institutions

Data backup



Manual

Copying files to relevant folders



Automated

- Install software
e.g. Time machine
(Mac users)

- RAID technology
- Checksums



Copying files to relevant folders

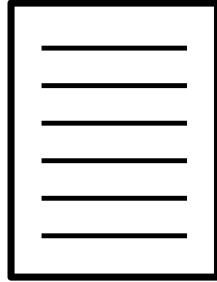
Automatically upload files to the cloud when any changes are saved

If manual ... how often?



How much would you be willing to lose?

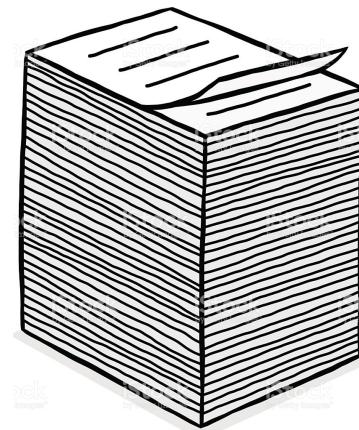
1 day



1 week



1 month-year



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

Data backup and file sharing



Dropbox



Google Drive



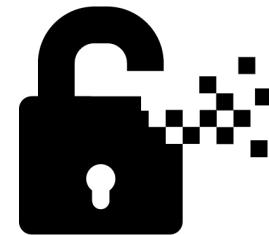
OneDrive

Space/price	2 GB (free) 1 TB (£96/year)	15 GB (free) 2 TB (£80/year)	5 GB (free) 1 TB (£55)
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?



CC-BY Danny Kingsley & Sarah Brown

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.

The screenshot shows the 'Data Management Guide' page of the University of Cambridge website. The header includes the university logo, navigation links for 'Study at Cambridge', 'About the University', 'Research at Cambridge', and 'Quick links'. Below the header is a main menu with 'Research Data' as the active item, showing a circular icon with a gear and a bar chart. The main content area is titled 'Data Management Guide' and features a photograph of hands typing on a keyboard. A sidebar on the left contains links for 'Research Data Management', 'Data Management Guide' (with sub-links for creating, organizing, and sharing data), 'Support', 'Data Repository', 'Data Policies', 'FAQ', 'News', 'Data Champions', 'Events', and 'Contact Us'. A right-hand sidebar lists 'Related links' for training workshops and depositing data, and 'Events' for workshops on post-publication sharing, FAIR data, and research support staff. At the bottom, there's a link to the 'Open Research Newsletter sign-up'.

What data to share?

- Data & metadata
 - Raw
 - Processed
- Code (software and scripts)
- Methods
- Papers
 - results and figures
- What about non-positive results?
 - RIO - Research Ideas and Outcomes



<https://riojournal.com/>

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



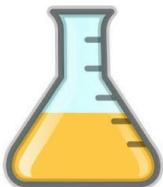
<https://www.biorxiv.org/>

DOAJ DIRECTORY OF
OPEN ACCESS
JOURNALS

<https://doaj.org/>

When to share?

Close 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 **Electronic Lab Notebook**



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/>

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.



UNIVERSITY OF
CAMBRIDGE

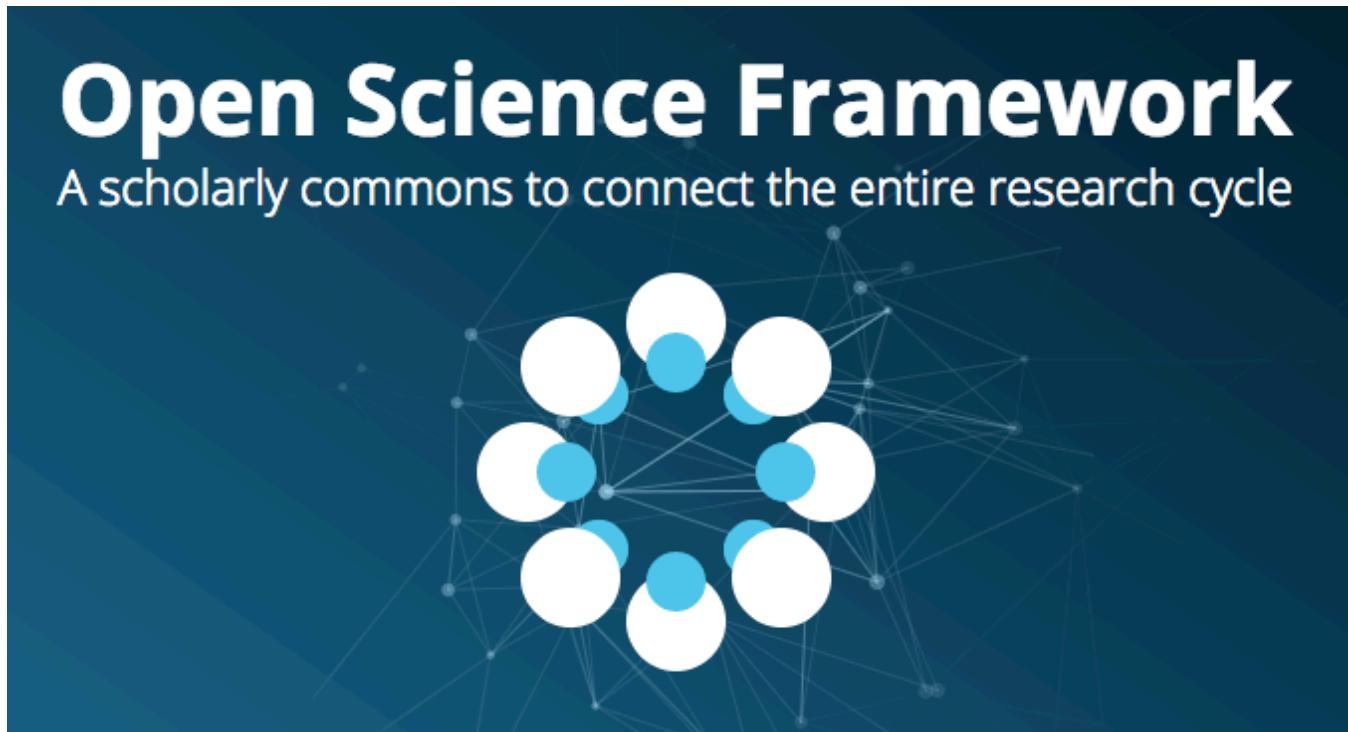
Accepted for publication?

Upload manuscript

<https://www.openaccess.cam.ac.uk/>

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/>

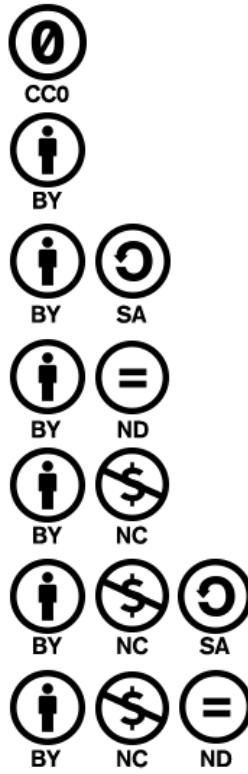


Choose a license

This chooser helps you determine which Creative Commons License is right for you in a few easy steps. If you are new to Creative Commons, you may also want to read Licensing Considerations before you [get started](#).



MOST OPEN



Share your code with one of the open source licenses

<https://opensource.org/licenses>

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



I want it simple and permissive.



open source
initiative
Approved License®

The [MIT License](#) is short and to the point. It lets people do almost anything they want with your project, including to make and distribute closed source versions.



I care about sharing improvements.

The [GNU GPLv3](#) also lets people do almost anything they want with your project, except to distribute closed source versions.

<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](https://www.cancerresearchuk.org/sites/default/files/policy_on_open_access.pdf)

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)



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**

