Welcome to The Carpentries Etherpad!

This pad is synchronized as you type, so that everyone viewing this page sees the same text. This allows you to collaborate seamlessly on documents.

Use of this service is restricted to members of The Carpentries community; this is not for general purpose use (for that, try [https://etherpad.wikimedia.org](https://etherpad.wikimedia.org/)).

Users are expected to follow our code of conduct: <https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html>

All content is publicly available under the Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>

 ----------------------------------------------------------------------------

**Dealing with data**

**Workshop led by BioRDM team on 6th November 2024, 9:30-12:30.**

**List of attendees (Please add your names and UUN!)**

-18 participants

 Please ALSO confirm you attendance here: <https://forms.office.com/e/Pb0Dx6H4YL>

**Open Science and FAIR principles**

**Exercise 1**

Below are some personal benefits to adopting Open Science practices. Read through them, select the 3 most important/attractive for you and mark them with +1, select two least important for you and mark them with 0

·         get extra value from your work (e.g. collaborators, reuse by modellers, ML specialists):+1+1+1+1+1+1+1+1+1+10+1+1

·         complying with funders’ policies:+1+1+1 +1+10

·         receive higher citations:000000000

·         demonstrate research impact:+1 +1+1+1+1+1+1+1

·         save own time (reproducibility but also communication overhead):+1 +1+1+1+1+1+1+1+1+1+1+1

·         become pioneers: +10

·         distinguish yourself from the crowd:000000000

·         plan successful research proposals:+1

·         gain valuable experience:+1+1+1

·         form community:0000+1

·         increased speed and/or ease of writing papers:+1+1 0

·         speed up and help with peer review: +10

·         build reputation and presence in the science community: +1 00+1+1

·         evidence of your scientific rigour and work ethic:+1 +1+1+1+1+1+1

·         avoid embarrassment/disaster when you cannot reproduce your results: +10+10

 DONE: +1+1+1+1+1+1+1+1+1++11+1+1+1+1+1+1

**Exercise 2**

**Data from publications**

(5+3 min)

**Exercise 2a. Impossible protocol (Room1, Room2)**

You need to do a western blot of the protein Titin, the largest protein in the body with a molecular weight of 3,800 kDa. You found a Titin-specific antibody sold by Sigma Aldrich (‘SAB1400284’) that has been validated in western blots and immunofluorescence. The Sigma SAB1400284 webpage lists the publication by Yu et al 2019 (<https://doi.org/10.1002/acn3.50831>) which uses the antibody.

**Can you find a complete protocol for separation and transfer of this large protein?**

·         Hint 1: Find the Western blot in the methods section.

·         Hint 2: Follow the references

How easy was it?

Answers:

- Reference 17 refers to standard protocol and doesn't mention details of the protocol

-Protocol was described by Avila et al (2014). It was easy to find. Reference 17.

-Fairly easy, protocol was in reference 17 of linked paper

- Referred to reference 17 for full protocol - western blotting protocol is not detailed though.

 - Reference 17 says just standard protocol

 - reference 17 is inaccessible

 is inaccessible

**Exercise 2b. Impossible average (Room3, Room4, Room5)**

The Ikram 2014 (<https://doi.org/10.1093/jxb/err244>) paper contains data about various metabolites in different accessions (genotypes) of *Arabidopsis plants.* You would like to calculate the average nitrogen content in plants grown under normal and nitrogen limited conditions.

**Please calculate the average (across genotypes) nitrogen content for both experimental conditions.**

* Hint 1. Data are in Supplementary data (Experiment 2 - <https://academic.oup.com/jxb/article/63/1/91/552676#supplementary-data> )
* Hint 2. Search for nitrogen in paper text to identify the correct data column.

Answers:

- Answer not obtained - Complexity of copy/pasting from a pdf, unclear if data has to be combined in a way the authors didn't initially foresee (combining root and shoot Nitrogen content)

-Difficulty in locating the correct data from the supplementary material

-  Relevant data is difficult to find. Excel file would have been easier for processing

- text file or excel file would have been much clearer to deliver the data

- Hard to locate data to calculate the average

- not proper description in supplementary

Exercise 3 10:16

(5+3 min)

Look at the dataset from Zenodo   <https://doi.org/10.5281/zenodo.6339631>

Identify elements that make this dataset FAIR

**Findable:**

**- Clear identifiers+1**

- Readme file +1+1+1+1+1+1

- Data Records in original publication links to two sources, including zenodo+1

- doi included+1+1+1

- Associated keywords

**Accessible:**

**- Includes view and download tracker**

- a way to easily download +1

-can download the entire dataset as a zip file+1

**Interoperable:**

**-.csv used for numerical data+1+1**

-multiple file formats provided as well as code to regenerate graphs in R

- github link

**Reusable:**

**-Standard data formats available**

-Code included for data processing and figure generation (no environment specifications though)+1

- Clear definitions of transformations used to produce each column+1

-License file

DONE: +1

--------------------------------------------

Break - back at 10:40

**Metadata**

**Exercise 4**

  (3min)

What information – metadata would you need to re-use the data like in example picture.

**Think as a consumer** of the data not producer.

Type your proposals:

- experimental procedures +1

* - transgene information
* - fluorescent protein

- Instrument used to measure +1 +1

- magnification / scale+1+1

- Capture parameters (exposure, time)+1

-Facilities used/location data recorded

- protocol of sample preparation (endogenous fluorescence or immunostaining)  +1 +1++1+1+1

-What is fluorescense indicating +1+1+1

- Statistical tests used/number of replicates

-experimental conditions (temperature, pH, etc.)aseptic conditions+1 +1+1+1

- species, strain +1 +1+1

- An image of the reference control +1

- Experimental conditions

--------------------------------------

**Record keeping**

**Exercise 5:**

(5 +3 min)

**Differences between analog and digital record keeping**

Compare the electronic version of the tea protocol:

<https://www.protocols.io/view/how-to-make-a-cup-of-tea-buhknt4w>

with the paper one from the photo:

<https://github.com/carpentries-incubator/fair-bio-practice/blob/gh-pages/fig/06-handwritten-tea-protocol.jpg>

What are advantages and disadvantages of traditional analog records vs digital records? Try to find at least a handful of advantages and disadvantages for each. With all of these, which system do you think is most advantageous?

**Room1 & Room2 & Room3 room**

Advantages of traditional analog records

-can record/add things on spot in the lab or during fieldwork+1+1+1

  - easier to carry a notebook than a laptop+1+1 (ipad :)

- don't need internet access +1+1

- Cannot be priced out of analog records

- You can usually see all the changes made (harder to manipulate)+1

Advantages of digital records

-Cleaner and easier to interpret

-Can usually do calculations in digital records (benchling has spreadsheet functionality)+1

-Can keep a dedicated laptop for use in-lab

- Much harder to lose

-easier to edit without making it look messy

- Accessible to multiple users at once/easily shareable

- Link to registry in lab

- TImestamped inputs on some digital lab books, better for patenting etc.

- indexing / contents - can search for content+1+1

- easier to read (than someone' s handwriting)+1+1

**Room4 & Room5 room**

Disadvantages of traditional analog records

-Might not be written in a tidy manner

-Difficulty in going back and finding the particular protocol as there are many experiments done after that

- Can be easily lost if single "copy" of the original record exists

- Harder to share

- we depend on good hand writing

Disadvantages of digital records

- History might not be included,

-May be lost if not properly backed up

- Knowledge needed before starting record keeping - how to version, backup, compatibility

- May not be accessible while executing the protocol

- need to finish experiments to record digitally (i am not supportor for this idea)(same?)

**LINKS**

Example record:

<https://benchling.com/s/etr-0FdV1H0rpWeHk4H72NOg/edit>

Our ELN resources

<https://www.wiki.ed.ac.uk/x/f0SkGw>

Benchilng tutorial:

<https://www.wiki.ed.ac.uk/display/RDMS/Benchling+%28quick%29+tutorial>

--------------------------------------------

**Working with Files**

**Exercise 6**

**A good name**

Select which file options adhere the best to the presented recommendations:

1.

a) analysis-20210906.xlsx +1+1+1+1+1+1+1

b) rna-levels-by-site.v002.xlsx+1+1+1+1+1+1+1+1+1

c) analysis of rna levels from 5Aug2021.xlsx+1+1

2.

a) 20210906-birds-count-EDI.csv+1+1+1+1+1+1++11+1+1+1+1+1+1+1+1+1+1+1

b) birds.csv

c) birds-count&diversity EDI 2021-09-06.csv

3.

a) 2020-7-12\_s2\_phyB\_+\_SD\_t01.raw.xlsx +1+1+1+1

b) ld\_phyA\_on\_s02-t01\_2020-07-12.norm.xlsx+1+1+1+1+1+1+1+1+1+1+1+1+1+1

c) ld\_phya\_ons\_02-01\_2020-07-12.norm.xlsx

DONE:+1+1+1+1+1+1+1+1+1+1+1+1+1+1

**Exercise 7 11:40**

**Projects structure**

Have a look at the four different folder structures A-D.

<https://github.com/carpentries-incubator/fair-bio-practice/blob/gh-pages/fig/07-file_organisation.png>

The first two” A) B) are recommended for computing, the other two: C) D) are for more wet/biological projects.

**Room1 & Room2:**

When/why would you use A) and when/why B)

A) when writing a manuscript/ thesis+1

B) Good for a long term project with explicit stages. Or if each stage is done by different collaborators+1+1

 - I would use it as organisation for project in general, so topic-related data are all together

**Room3 & Room4 & Room5:**

When/why would you use C) and when/why D)

C) Project C may be more useful in contextualizing overall behaviour of an individual

- C is probably more useful at the later stages of a project where specific results have been derived from the raw data +1

- maybe better for a smaller sample size

D) Project D may be more useful in contextualizing overall usefulness of a procedure / relevance of timepoint

- D has filenames with descriptors of where the sample came from, better for picking out from a large data set

 -D seems to have more data driven sets and the analysis might have to be done as the result section is not there.

DONE:

----------------------------------------------

**Resources for Data Management**

Course materials:

<https://biordm.github.io/SBS-PhD-Induction-Dealing_with_data/>

BioRDM wiki

<https://www.wiki.ed.ac.uk/display/RDMS>

RDS page

<https://www.ed.ac.uk/information-services/research-support/research-data-service>

DRS page

<https://digitalresearchservices.ed.ac.uk/>

DataStore

•       <https://www.ed.ac.uk/information-services/computing/desktop-personal/connect-uni-file-storage>

•       <https://www.wiki.ed.ac.uk/x/tet_H>

Sharepoint

[https://uoe-my.sharepoint.com](https://uoe-my.sharepoint.com/)

UoE WIKI

[https://www.wiki.ed.ac.uk](https://www.wiki.ed.ac.uk/)

DMPOnline

[https://dmponline.dcc.ac.uk](https://dmponline.dcc.ac.uk/)

Jupyter notebook

<https://jupyter.org/>

Benchling

<https://www.benchling.com/>

Protocols.io

<https://www.protocols.io/>

Zenodo

<https://zenodo.org/>

UoE DataShare

<https://datashare.is.ed.ac.uk/>

FAIRification of you project (FAIR cookbook)

<https://faircookbook.elixir-europe.org/>

FAIR in (bio) practice

<https://carpentries-incubator.github.io/fair-bio-practice/>

 Github/ GitLab

<https://github.com/>

<https://git.ecdf.ed.ac.uk/>

Version control with Git - self paced course

<https://swcarpentry.github.io/git-novice/>

**Exercise 8**

**Quiz**

Which of the following statements are true/false? T or F

·         F in FAIR stands for free. FFFFFFFFFFFFFFF

·         Sharing numerical data as a .pdf in Zenodo is FAIR. FFFFFFFFFFFFFFF

·         Sharing data as an Excel file is not FAIR. FFFFFFFTFFFfFFFF

·         Group website is a good place to share your data. FFFFFFFFFFFFFFFF

·         Data from failed experiments are not re-usable. FFFFFFFFFFFFFFF

·         Data should always be converted to Excel or .csv files in order to be FAIR. FFFFFFFFFTFFF

·         A DOI of a dataset helps in getting credit. TTTTTTTTTTTTTTT

·         FAIR data are peer reviewed. FFFFFFFFFFFF

·         Open Science relies strongly on the internetTTTTTTTTTTTTT

·         Good record keeping ensures transparencyT TTTTTTTTTTTT

·         There are advantages to using analog record keeping when compared to digital record keeping. TTTTTTTTTTTTTT

·         On balance, digital record keeping is more advantageous than analog record keeping. TTTTTTYTTTTTT

·         ‘output 3-Aug-2022’ is a good file name FFFFFFFFFFFFFFF

·         Digital records are easier to search (for and within) than analog records. TTTTTTTT TTTTTT

DONE:+1+1+1++1+1+1+1

-----------------------------------------------------------------------------------------------------------------

**Feedback:**

1.      How do you feel about the presented topics after this session (type +1 next to the statement that best describes your feeling):

•       I am more confused:

•       I have a better understanding of them now: +1+1+1+1++1+1+1+1+1+1+1+1+1+1+1

•       My knowledge has not changed much: +1

2.      How was the pace of the lesson:

•       Too fast:

•       About right:+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1

•       Too slow: +1

3. If the lesson could be 5 minutes longer, what would you add or spend more time on:

I would have hoped for more concrete examples of best practices +1+1+1

 The existing data sharing repositories and what they are for+1+1

 FAIR practices for code (as opposed to data; e.g., requirements, version control, ...)

 Maybe tips from current PhD students would have been helpful

4. What could be improved:

 Seems like many groups use benchling, maybe a section on specific benchling functions would be useful+1+1

 Concrete practices/examples for data backup - You must use DataStore

5. What did you like:

Introduction to Zenodo, use of real data+1