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/>

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

**FAIR in (Biological) Practice**

**DAY SECOND PAD at**[**https://pad.carpentries.org/2022-02-16-ed-dash-fair**](https://pad.carpentries.org/2022-02-16-ed-dash-fair)

[**https://edcarp.github.io/2022-02-15\_ed-dash\_fair-bio-practice/**](https://edcarp.github.io/2022-02-15_ed-dash_fair-bio-practice/)

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

**Important notice:**

Before you begin today, please take some time to sign up for the following two accounts:

1. Benchling     (the ELN we will use for one of lessons): [https://benchling.com](https://benchling.com/).
2. Protocols.io     (the protocol repository with PID, which we will also be using for today's     lessons): <https://www.protocols.io/>

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

 13:10

**List of attendees**

- Kadi Vaher

-Alexandre Meier

-Flavia Fonseca Pezzini

-Wu Huang

-Ines Boehm

-Amelia Edmondson-Stait

-Adelaide Young

-Caity Ellis

-Livia Scorza

-Marta Amador

-T Zhou

-Sumy Vnnilathil Baby

-Nneka Nnadi

- Cammy Beyts

-Cigdem Selli

1. **You and data sharing**

Thinking of how you make your data or code available to others and how you use others data, write +1 next to the statments that matches your own experience:

- I do not really share data, I only publish the results as a part of a publication:

- I have made my data available only as Supporting Information for a paper:

- I have made my data available as both Supporting Information and as a dataset in a repository: +1+1+1+1+1+1

- I have made my data/code available without having it published in a paper: +1+1+1

-  I share my code in github or another code repository: +1+1+1+1+1+1+1+1+1

- I make my code available on demand:+1+1+1

- I have used a dataset from a public repository:+1+1+1+1+1+1+1+1

- I have used others code from github or such:+1 +1+1+1+1+1+1+1+1+1

DONE: +1+1+1

**2. You and your colegues 3+0.5, 13:15**

 In your group, introduce yourself and your background.

 Discuss what are your reason to attend the course.

 Green

 - backgrounds: postdoc, phd students bioinformatics/biological sciences

 - reasons how to publish and share data

 Blue

 - backgrounds: neuroscience, medicine;, developmental biology; PhD students, honours student

 - reasons: don't have ideas about open science and look for new input; to use time productively; honours project is open science

 Red

 - backgrounds: genomics, cancer research, psychiatry, clinical research

 - reasons: reusable pipelines for publications, better data management and computational workflows, skill required by pharma industry

 Yellow

 - backgrounds: bioinformatics, plant molecular bio, stem cell biology

 - reasons: help others researchers (and themselves) in the group in sharing data using FAIR practices

 13:21

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

**Lesson 2: Open ScienceExercise 1. Benefits of openness 13:32 (4.5)**

Being open has other benefits beyond giving free access to information.

For example “Open Access”:

·       speed of work and knowledge distribution

·       new metrics of impact: views, downloads, tweets etc

Discuss in your group additional benefits, or addressed problems for the selected open practices, type them bellow:

(Green Room) Open Data:

-Ability to cross compare/validate data, reuse data

-Timing of data sharing is important

-Making the methodologies used transparent and reusable for saving time

(Blue Room) Open Software: analysis, visualisation; don't have to invent the wheel and write scripts from scratch; learning tool - e.g. if don't have too much coding experience

BUT might use implementation that is not actually correct for their data; could be time delay; need to make sure you are cited/accredited for your work; needs to be maintained, high risk for obsolete code.

-

-

(Red Room) Open Notebooks:

- for other people to replicate your code/workflow

- collaborators/other people in the field get to know you

- tracking who is using your code, eg. forking on github

(Yellow Room) Open Peer Review:

- if you know the reviewer it might be easier to understand what they want

-avoid conflict of interest

-disadvantage: a reviewer might be afraid of backlash when they believe the paper shouldn't be published and the authors are "more senior" in the field

-can lead to collaborations

(Orange) Open Educational materials:

-

-

DONE:

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

**Exercise 2. Personal benefits of being “open” 13:46**

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

·       complying with funders’ policies:00

·       receive higher citations:00+10

·       demonstrate research impact:+1

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

·       become pioneers:000

·       distinguish yourself from the crowd:0000

·       plan successful research proposals:+10+1

·       gain valuable experience:+1

·       form community:+1

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

·       speed up and help with peer review:+1+1+1

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

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

·       avoid embarrassment/disaster when you cannot reproduce your results: 00000

 DONE: +1+1

 13:52

Can you think of other benefits? How do personal benefits of Open Science compare to the benefits for the (scientific) society.

**OA links**

Details of funding bodies and their involvement and requirements can be found at

Plan S/cOAlition S: <https://www.coalition-s.org/plan-s-funders-implementation/>

There is also a cOAlition S journal checker tool (<https://www.coalition-s.org/blog/unboxing-the-journal-checker-tool/>) to assess compliance being developed. The Directory of Open Access Journals (DOAJ - <https://doaj.org/>) is a tool to find which journals are Open Access.

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

**Exercise 3. Why we are not doing Open Science already 13:52**

Discuss Open Science barriers, type bellow the reasons for not being open:

-competition between research groups +1

-working with clinical data that needs permission for being shared +1 (difficult to reconsent patients - difficulty working between healthboards Eng and Scot...)

- Don't know how to do it

- Not being asked to do it

-Not having the time

- investing in the time to learn the skills +1

 -seems intimidating

 - many steps to think of - don't want to do anything wrong

- embarrassed of code quality/readability +1+1+1

Too many different pipelines available, no idea which is the best one +1

-Not knowing what platforms are the best for sharing data - I know github but what about other data like westerns/qPCR etc?

differnet people share their code/analysis pipeline at different levels in different details

Too many code files to share if we want to share all....

- so many repositories!!+1

Where to next links

•  Challenges & benefits of OS: <https://doi.org/10.1371/journal.pbio.3000246>

•  Centre for Open Science: <https://www.cos.io/>

•  Ted talk supporting OS: <https://youtu.be/c-bemNZ-IqA>

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

**Exercise 4. Open Science Quiz 2:04**

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

* ·       Open Science relies strongly on the internet:TTT TTTTTTTT

·       Open Access eliminates publishing costs:FFFFFFFFFFFFF

·       Open Data facilitates re-use: TTTTTTTTTTTTT

·       Open Data increases confidence in research findings:TTTTTTTTTTTT

·       In Open Peer Review, readers vote on publication acceptance:FFFFFFFFFFFF

·       Open Notebooks improve reproducibility:TTTTTTTTTTTTT

·       Open Notebooks can create patenting issues:TTTTTTFTTT

·       Open Access permits the whole society to benefit from scientific findings: TTTTTTTTTTTTT

·       Citizen Science engages public in the research process:TTTTTTTTTTT

·       Citizen Science can help getting  ecological data quickly:TTTTTTTTTTT

DONE: +1+1

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

**Lesson 3: Being FAIR 14:15**

**Exercise 1a. Protocol (green, blue) 4.5**

**14:21**

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 an antibody sold by Sigma Aldrich that has been validated in western blots and immunofluorescence. Sigma Aldrich lists the publication by Yu et al 2019 (<https://doi.org/10.1002/acn3.50831>) which uses their 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?

BLUE: the 2019 paper references this:

<https://doi.org/10.1002/ana.24102> which says that Western plotting was "performed according to standard methods" WELL DONE BLUE

**Exercise 1b. Average content (red, yellow)**

The Ikram 2014 (<https://doi.org/10.1093/jxb/err244>) paper contains data about various metabolites in different accessions (genotypes) of *Arabidopsis plant.* 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://tinyurl.com/hjkdzsd4>)

·       Hint 2. Search for nitrogen in paper text to identify the correct data column.

-issues with copying from pdf; legends are hard to find and correlate

-acronyms of data columns not annotated!

DONE:

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

**Exercise 2. FAIR Example 12:54**

Uniprot is a high-quality and freely accessible resource of protein sequence and functional information. Have a look at the record of the GFP protein: <https://www.uniprot.org/uniprot/P42212>

**Identify elements that make it FAIR**

Findable:permanent unique identifier+1+1+1

 Names & Taxonomy [general scientific names]+1

Accessible:

 open for download

Interoperable: can download the protein sequence in different formats (e.g. FASTA)+1, Links to papers and other databases to provide more information about the protein. use of ontologies and standard vocabulary Links to other data repositories

Reusable reviewed by a dedicated data curator (not the person generating the data/metadata), a lot of metdata and info about this protein+1

DONE:

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

**Exercise 3. FAIR and You 15:2 (4)**

The FAIR acronym is sometimes accompanied with the following labels:

·       Findable - Citable:

- gives us more citations

- facilitates collaborations by being easier to find and visible

-Citation helps to find the data easier

·       Accessible - Trackable and countable

- also facilitate collaborations - easier to access each others data/methods/workflows/code etc

- easier to access others work, identify collaborators

-allows gathering of metrics so we can identify people who are interested in our work, could encourage communication to different groups who you wouldnt think would be interested in your research

·       Interoperable - Intelligible

-helps organise the data in a way that is useful for myself in the future

- if you have good metadata (intelligible), less people asking you questions

-Data needs to be legible for other tems to operate on it

·       Reusable - Reproducible

- By knowing what others have done you can replicate experiments exactly making science hopefully more reproducible/reliable

- other researchers can reuse your dataset - able to track and credit the original data

The future me will be able to use it again

-  Data need to be in proper format and commented so that other people can go through same steps and reproduce it

- others after me will be able to just step into my position quicker rather than requiring in-depth inductions in all procedures if all my code, workflows etc are available

Using those labels as hints discuss how FAIR principles directly benefit you as the data creators.

DONE:

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

**Exercise 4. FAIR Quiz 3:15**

Which of the following statements is true/false (T or F).

·       F in FAIR stands for free. FFFFFFFFFFFF

·       Only figures presenting results of statistical analysis need underlying numerical data. FFFFFFFFFFFF

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

·       Sharing numerical data as an Excel file via Github is not FAIR. FFFFFFFFFFFF

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

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

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

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

·       FAIR data are peer reviewed. FFFFFFFFFFF

·       FAIR data accompany a publication. FTFFFFFFF

DONE:+1

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

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

BACK 3:25

**Lesson 5: Intro to metadata 3:30**

**Exercise 1. Identify types of metadata 3:42 3.5**

Here we have an excel spreadsheet that contains project metadata for a made-up experiment of plant metabolites: <https://carpentries-incubator.github.io/fair-bio-practice/fig/04-metadatafull_spreadsheet.png>

In groups, identify different types of metadata (administrative, descriptive, structural) present in this example.

Just as a reminder:

•Administrative: relevant to managing it

  e.g. Experimental code, PI

  - funder/grant number

  Contact

  -PI

•Descriptive/citation: assists with discovery/identity

  e.g. Authors, persistent identifier

  - study title and dates

  date range

 Title

 -Author

•Structural: how the data came about & is structured

  e.g. Collection method, folder structures

 - abbreviation explanations

 - measurement units in column names

 - columns containing experimental conditions

-protocols

-Growth and measurement protocols

DONE:

 Examples of metadata can be data:

 Genotypes/Strains used

 light conditions

 media used

 cell counts in a IHC image

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

**Minima Information Standards**

<https://fairsharing.org/collection/MIBBI>

<https://fairsharing.org/standards/>

**Exercise 2. Minimal Information Standard**

Look at Minimum Information about Neuroscience Investigation (MINI) Electrophysiology

<https://www.nature.com/articles/npre.2008.1720.1.pdf>

which contains recommendations for reporting the use of electrophysiology in a neuroscience study. (Neuroscience, or neurobiology, is the scientific study of the nervous system)

Scroll to **Reporting requirement** and decide which of the points 1-8 are:

**a)** important for understanding and reuse of data: 2 212, 3, 4, 5, 8,8

 1c,4, 7

**b)** important for technical replication:

 2,4,33,54, 3, 7,6567, 3,7

**c)** could be applied to other experiments in neuroscience:

 4, 6, 5, 7, 8, 2, 3

DONE:

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

**Exercise 3. What to include – discussion 4:04**

Think of the data you generate in your projects and imagine you are going to share them.

What information would another researcher need to understand or reproduce your data (the structural metadata)?

**Think as a consumer** of your data not the producer!

For example, we believe that any dataset should have:

·       A name/title

·       Experiment purpose or experimental hypothesis

Write down your proposals:

 the statistics tests used and parameters +1

 date and time of the data being produced

 animal M/F Age

 version of the software used+1

 genetic mutations (how were mutations introduced), what species

 experimental factors, buffers, drugs used, controls

 plasmids used

 definition of variables (to avoid the problem with nitrogen example)+1+1

 LOT numbers (and batch numbers)+1

 ethical approvals+1

 Methodologies used toderive the data (incl detailed protocols)

 and why changes were made in methods if they were made

 class definitions

 packages used (R) + version nr

 Model used eg cell line/bacterial strain

 basic info on study objects (e.g. age, sex etc)

 summary statistics, how much missing data there is

DONE:

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

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

**Lesson 4: Intellectual property, Licencing and Openness**

**Exercise 1. Checking common licence:**

1. Open CC BY licence summary <https://creativecommons.org/licenses/by/4.0/>

is it clear how you can use the data under this licence and why it is popular in academia?

2. Check the MIT licence wording: <https://opensource.org/licenses/MIT>

is it clear what you can do with software code under this licence?

3. Compare the full wording of CC BY

<https://creativecommons.org/licenses/by/4.0/legalcode>

can you guess why the MIT licence is currently the most popular for open source code?

DONE:

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

**Lesson 6: Being precise 4:15**

*If you have not done it yet, register yourself on ORCID (*[*https://orcid.org/*](https://orcid.org/)*)*

**Exercise 1. Public ID in action 1**

The Wellcome Open Research journal uses ORCID to identify authors.

Open one of our papers <https://doi.org/10.12688/wellcomeopenres.15341.2> and check how public IDs such as ORCID can be used to interlink information.

-

-

-

-

DONE:

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

**Registries**

•species e.g. NCBI taxonomy

<https://www.ncbi.nlm.nih.gov/Taxonomy>

•chemicals e.g. ChEBI

<https://www.ebi.ac.uk/chebi>

•proteins e.g. UniProt

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

•genes e.g. GenBank

<https://www.ncbi.nlm.nih.gov/genbank/>

•metabolic reactions, enzymes e.g KEGG

<https://www.genome.jp/kegg/>

**Exercise 2. Public ID in action 2 4:20**

The second metadata example (the Excel table) contains two other types of public IDs.

<https://carpentries-incubator.github.io/fair-bio-practice/fig/04-metadatafull_spreadsheet.png>

 ADD LINKS TO EXCELL FILES AS WELL!!!!!!!

·       Can you find them?

 B5 B2

 Kegg IDs+1

 Tair ID?

·       Can you find the meaning behind those IDs?

-

-Gene Locus

DONE:

**Q&A:**

Do you have any questions about the topics dicussed today? Please write them down here. Use +1 to upvote the ones you are interested in if someone already asked it. We will briefly discuss them before the following set of lessons.

-

-

-

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

**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

•       My knowledge has not changed much:

2.      How was the pace of the lesson:

•       Too fast:+1 (IP)

•       About right: +1+1+1+1+1+1

•       Too slow:+1 (for the begining part open science)

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

 -slightly longer discussions during the exercises done in breakout rooms+1+1+1

 coffee break

4. What could be improved:

- not sure the breakout rooms are working well for discussing the topics/excercises

- the above probably because there isn't enough time? or maybe need facilitators to the breakoutrooms - I think its a balance some we could;ve talked for longer others it was probably the right amount of time - thats probably not very helpful!

5. What did you like:

- I like the excercises+1+1

 I like use of pad and format in general+1+1+1+1