# FAIR in (Biological) Practice

## DAY SECOND PAD at

## <https://pad.carpentries.org/2021-10-20_ed-dash_fair-bio-practice>

### <https://edcarp.github.io/2021-10-19_ed-dash_fair-bio-practice/>

Online, 19 - 22 October 2021, 13:00 - 17:00

## Day 1 - Tuesday 19 October

### List of attendees

#### 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:

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

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

- I make my code available on demand:

- I have used a dataset from a public repository:

- I have used others code from github or such:

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

### Lesson 2: Open Science

#### Exercise 1. Benefits of openness

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:

-

-

(Blue Room) Open Software:

-

-

(Red Room) Open Notebooks:

-

-

(Yellow Room) Open Peer Review:

-

-

-

(Orange) Open Educational materials:

-

-

DONE:

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

#### Exercise 2. Personal benefits of being “open”

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):
* complying with funders’ policies:
* receive higher citations:
* demonstrate research impact:
* save own time (reproducibility but also communication overhead):
* become pioneers:
* distinguish yourself from the crowd:
* plan successful research proposals:
* gain valuable experience:
* form community:
* increased speed and/or ease of writing papers:
* speed up and help with peer review:
* build reputation and presence in the science community:
* evidence of your scientific rigour and work ethic:
* avoid embarrassment/disaster when you cannot reproduce your results:

 DONE:

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

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

-

-

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

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

* Open Science relies strongly on the internet:
* Open Access eliminates publishing costs:
* Open Data facilitates re-use:
* Open Data increases confidence in research findings:
* In Open Peer Review, readers vote on publication acceptance:
* Open Notebooks improve reproducibility:
* Open Notebooks can create patenting issues:
* Open Access permits the whole society to benefit from scientific findings:
* Citizen Science engages public in the research process:
* Citizen Science can help getting  ecological data quickly:

DONE:

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

### Lesson 3: Being FAIR

#### Exercise 1a. Protocol (green, blue)

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?

#### 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.

**Exercise 1c. Data from figure**

Systems biologists usually require raw numerical data to build their models. However, those are sometimes not easy to find. Take a look at the following example:

Try to **find the numerical data** behind the graph shown in **Figure 6** (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC166576/figure/F6/>) which demonstrates changes in levels of phytochrome proteins of Sharrock RA and Clack T, 2002 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC166576/>).

* Hint 1: Materials and methods describe the quantification procedure
* Hint 2: Supporting Information or supplementary materials sections often contain data files.

How easy was it?

-

-

DONE:

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

#### Exercise 2. FAIR Example

Zenodo is general data repository.

Have a look at the dataset record with COVID-19 data:

<https://doi.org/10.5281/zenodo.6339631>

*Hint: navigate to linked github record to easily access the README file*

**Identify elements that make it FAIR**

Findable:

Accessible

Interoperable

Reusable

DONE:

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

#### Exercise 3. FAIR and You

The FAIR acronym is sometimes accompanied with the following labels:

* Findable - Citable
* Accessible - Trackable and countable
* Interoperable - Intelligible
* Reusable - Reproducible

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

DONE:

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

#### Exercise 4. FAIR Quiz

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

* F in FAIR stands for free.
* Only figures presenting results of statistical analysis need underlying numerical data.
* Sharing numerical data as a .pdf in Zenodo is FAIR.
* Sharing numerical data as an Excel file via Github is not FAIR.
* Group website is a good place to share your data.
* Data from failed experiments are not re-usable.
* Data should always be converted to Excel or .csv files in order to be FAIR.
* A DOI of a dataset helps in getting credit.
* FAIR data are peer reviewed.
* FAIR data accompany a publication.

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 one of the most popular for open source code?

DONE:

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

### Lesson 5: Intro to metadata

#### Exercise 1. Identify types of metadata

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

•Descriptive/citation: assists with discovery/identity

  e.g. Authors, persistent identifier

•Structural: how the data came about & is structured

  e.g. Collection method, folder structures

DONE:

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

**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:

**b)** important for technical replication:

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

DONE:

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

#### Exercise 3. What to include

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:

*Hint: Let’s start with the microscope image example*

-

-

-

DONE:

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

### Lesson 6: Being precise

*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

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>

* Can you find them?
* Can you find the meaning behind those IDs?

-

-

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:

•       My knowledge has not changed much:

2.      How was the pace of the lesson:

•       Too fast:

•       About right:

•       Too slow:

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

4. What could be improved:

5. What did you like: