# FAIR for busy biologists

### 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: Open Science and FAIR principles

#### Exercise 1: 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.

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

#### Exercise 2: Why we are not doing Open Science already

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>

### Being FAIR

#### Exercise 3a. 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 3b. 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://github.com/carpentries-incubator/fair-bio-practice/raw/gh-pages/files/err244_Supplementary_Data-2023-03-28.zip> )
* Hint 2. Search for nitrogen in paper text to identify the correct data column.

DONE:

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

#### Exercise 4. 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:

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

DONE:

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

#### Exercise 4. FAIR Quiz

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

* Open Science relies strongly on the Internet
* Open Access eliminates publishing costs
* Open Data facilitates re-use
* Open Data can increases confidence in research findings
* In Open Peer Review, readers vote on publication acceptance
* Open Access permits the whole society to benefit from scientific findings
* Citizen Science engages the public in the research process
* Release of public datasets is important for career progression
* 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 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 2: Intro to metadata

#### Exercise 1. 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:

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

**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:

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

### Lesson 3: Meta(data) in Excel

#### Exercise 1: What can go wrong with data in Excel

Have a look at the example excel data-file:

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

<https://github.com/carpentries-incubator/fair-bio-practice/raw/gh-pages/files/04-bad-metadata.xlsx>

**Questions:**

- What do you find confusing?

- What would you try to clarify with the author before doing anything with the file?

- What will be the issues with calculation of: average biomas, biomas per genotype?

- Typically, more advance data analysis is done programmatically, which requires e.g. conversion to a text format as csv, tsv format. Or using a library that reads Excel file and "kind of makes this conversion on the fly". Save this file in a text format, close Excel and reopen the saved files. What has changed?

**Answers:**

-

-

-

Have you seen similar tables? Do you believe this example is realistic? (add +1)

DONE:

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

#### Exercise 2: Spotting problems

Look at the following rows and columns in the problematic table:

·    Row 5

·    Row 2

·    Column C

·    Column E

·    Column L

Table image: <https://carpentries-incubator.github.io/fair-bio-practice/fig/bad-metadata.png>,  
Excel file: <https://github.com/carpentries-incubator/fair-bio-practice/raw/gh-pages/files/04-bad-metadata.xlsx>

which of the problems discussed before can you spot in these rows and columns.

Here, we list them again:

1.     Using multiple tables

2.     Using multiple tabs

3.     Not filling in zeros

4.     Using problematic null values

5.     Using formatting to convey information and organizing data

6.     Placing comments or units in cells

7.     Entering more than one piece of information in a cell

8.     Inconsistency in used values

9.     Using problematic field names

10.  Using special characters in data

11.  Values without field labels

Type the problem number(s) next to the table elements

·    Row 5:

·    Row 2:

·    Column C:

·    Column E:

·    Column L:

DONE:

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

#### Exercise 3: Outsmarted by Excel

Open Excel and type the following values into the cells:

A       B       C       D       E       F

Gene    Sept2   Sample  0013    Record  12/5/4

Mar/1   1March  Mar-1   1-3     14/3/20 43904

**Questions:**

* Is what you see what typed?
* Can you force the above formatting?
* Do you know which year these dates represent?

DONE:

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

#### Exercise 4: Data tables Quiz:

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

* Do’s and don’ts help in programmatic re-use:
* Avoiding multiple tabs improves interoperability:
* Having accompanying README file with a table description is not FAIR:
* No ‘spaces’ in columns headers improve readability:
* 2022-07-15 is ISO date format:
* 20220715 date format is better for excel than 2022-07-15:
* “No data” is better than leaving cell “blank” for missing data:

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.     On the scale 0 - 5 (zero a terrible lesson, 5 a fantastic lesson)

How good were the lessons:

-

2. On the scale 0 - 5 (zero not at all, 5 yes it was productive way of spending my time)

Was it worth your time:

3. 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:

4.     How was the pace of the lesson:

•       Too fast:

•       About right:

•       Too slow:

5. What could be improved:

-

-

-

-

6. What did you like:

-

-

-

-