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 for busy biologists**

**20/03/2024 - Day 1; 13:00 - 16:00**

Write your name in the assigned breakout room

**Number of participants: 19**

**You and data sharing 13:11 - 13:14**

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

- I do not really share data, I only publish the results as a part of a publication:+1+1+1+1+1

- I have made my data available only as Supporting Information for a paper:+1 +1+1

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

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

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

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

 - I support and advisors researchers on how to archive, share and publish data: +1+1+1

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

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

**Lesson: Open Science and FAIR principles 13:15**

**Exercise 1: Personal benefits of being “open” 13:25 13:30**

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+1+1+1+1+1

·         complying with funders’ policies: +100000+100+1

·         receive higher citations:000000000

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

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

·         become pioneers:+10+10

·         distinguish yourself from the crowd:+10000000+1

·         plan successful research proposals:+1+10+1+10

·         gain valuable experience+1:0+1+10

·         form community:+1+1+1+100

·         increased speed and/or ease of writing papers:0+1000+10

·         speed up and help with peer review:00000+1

·         build reputation and presence in the science community:++11+1++1+10+1

·         evidence of your scientific rigour and work ethic: +1+1+1++11+1+1+1+1++11+1+1+1

·         avoid embarrassment/disaster when you cannot reproduce your results:+100000

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

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

 Knowledge sharing

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

**Exercise 2: Why we are not doing Open Science already 13:36**

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

- Fear of plagiarism=1

- Infrastructure isn't in place+1+1 +1 (and infrustructire isn't easy to navigate nor quick to learn eg.github)

- Rules and regulations are complicated +1+1+1

-Too many other demands on my time

- Anxiety about balance between GDPR and open science+1+1

- Lack of knowledge of open data approaches +1 +1+1+1

- Don't know how+1

 - Takes a long time to make the data/code clear and reusable to others+1+1+1+1

- Lack of awareness of its impact+1+1+1+1+1+!

-Intellectual Property+1 - Pharma companies do not share their code/datasets.

-to distinguish yourself from the crowd

General lack of motivation to do something like that

- Fear of change

- Obstinate "This is how I've always done it and I've got my professorship so why should I change.."+1+1

-New habits to create

-Competition makes it easy for people to steal and publish your data before you

-Becuse it's not a REF requirement, it's not a priority

DONE: +1+1\+1+1+1+1+1+1+1+1+1+1+1+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>

**Being FAIR**

**Exercise 3a. Protocol (1,2) 13:49 - 5 minutes - 10 minutes talking**

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?

Followed the references and wasn't easy to find. Yu et al cites a different paper which is behind a paywall. I don't have access to the journal. But I found the paper on sci-hub. So, I do have a protocol.

Did you find out how to make the western?

easy to find the reference the paper uses but the methods of that paper mention several differnt antibodies (I think?)

 Paper refs another paper which details the protocol but it is behind a pay wall - Imperial College does not have institutional access.

 Quite easy if you follow the reference (but I have institutional access)

 Followed the trail to "standard methods" without a citation. Answer to Q is no.

Exercise 3b. Average content (3,4)

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:

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

**BREAK - back at 14:15**

**Exercise 4. FAIR Example 14:24 (5 minutes)**

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:

Clear data description+1+1+1+1

links to publication and supplements+1

Persistent links to data with different versions+1

Persistent Identifier (DOI)+1+1+1

Accessible

 Clearly labelled and easy to navigate the data sets+1+1

 Publicly accessible and clearly labelled with metadata+!

 Open licence (CC-BY)+1

Interoperable

 provides files and code in common format (csv, R)+1+1+1

 no bespoke file formats (not sure about .R or .xcf?)

 clear common formats

 csv - open / non-proprietory format+1

- Medatada

Reusable

 detailed Readme file+1+1+1+1

 all scripts included, especially data transformations +1

 all in the excel format which make easy to reuse it

Code and associated data are easily accessible +1

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

 14:32

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

DONE:

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

**Exercise 4. FAIR Quiz 14:34**

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

·         Open Science relies strongly on the InternetTTTTTTTTTTTTT

TTTTT

·         Open Access eliminates publishing costsFFFFFFFFFFFFFT/FFF

·         Open Data facilitates re-useTTTTTTTTTTTTTTTT

·         Open Data can increase confidence in research findingsTTTTTTTTTTTTTT

·         In Open Peer Review, readers vote on publication acceptanceFF  FFFFFFFFFFFF

·         Open Access permits the whole society to benefit from scientific findingsTTTTTTTTT(in theory) TTTTTT

·         Citizen Science engages the public in the research processTTTTTTTTTTTTTTT

·         Release of public datasets is important for career progressionFFFFTFFFTTT(if mandated by the funding agent)FF

·         F in FAIR stands for free.FFFFFFFFFFFFFF

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

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

·         Sharing numerical data as an Excel file via Github is not FAIR.FFFFFFFFFFFFFT(if xlsx but not .xls)

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

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

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

·         FAIR data are peer reviewed.FFFFFFFFFFFFFFF

·         FAIR data accompany a publication.TTTTT(not always)TTT(not always)SometimesSometimesTF

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

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

**Lesson 2: Intro to metadata 14:38**

**Exercise 1. What to include 14:51 14:53 (5 minutes) answers finished: 15:03**

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:

Room1: Example 1

(a plant transcriptomics data with different treatments; DEG: differentially expressed genes)

File: <https://github.com/tzielins/private-fair-for-busy-biologists/blob/main/episodes/fig/metadata_ex1.jpg>

-Attribution metadata Who? When? Where?

-Technical metadata relating to the methods

-Project purpose.

-

-

-

Room 2: Example 2 (insect repellent experiment)

File: <https://github.com/tzielins/private-fair-for-busy-biologists/blob/main/episodes/fig/metadata_ex2.jpg>

- Name & description of the experiment+1

- Description of the experiment setting+1

- Sample name/groups, technique, statistical analysis, D.mel cycle (experinmental condition)+1

- species strains/variants/source+1

- what time the experiements/assays were done/ were they done in blocks? +1+1

- Better annotation of experimental conditions - what is the %? etc +1

-Clear description of indices that are being calculated+1

-What is the test odor in panel A

Room3: Example 3 (protein interaction study; SUMO: Small Ubiquitin-like Modifier proteins; gel SDS-PAGE)

File: <https://github.com/tzielins/private-fair-for-busy-biologists/blob/main/episodes/fig/metadata_ex3.jpg>

- Legend for each image

- Detailed protocol for all figures

- Sample names explanation

- Tested Hypothesis

- Statistical test used

- Alignment software

- Protein database used

- Protein expression and purification

Room 4: Example 4 (circadian data - sleep pattern in flies )

File: <https://github.com/tzielins/private-fair-for-busy-biologists/blob/main/episodes/fig/metadata_ex4.jpg>

-Title and description of hypothesis

-Fly species, age, experimental conditions, number of flies

-statistics performed

-How dopamine and octopamine are quantified by FRET+1

-daily sleep time - how this is quantified, movement?

-how sleep time varies with age

DONE:+1

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

 Break, see you 15:10

**Minima Information Standards**

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

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

**Exercise 2. Minimal Information Standard 15:13 15:17 15:19**

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:

 1+1+1+1+1+1+1

 2+1+1

 3

 5+!+1+1

 6+1

 4+1

 7+1

 8+1+1+1+1+1+1

**b)** important for technical replication:

 1+1+1+!

 4

 2+1+1+1+1+1+1+1

 3+1+1+1+1+1

 4+1+1+1+1

 5+1+1+1+1

 6+1+1+1+1+1+1

 7+1+1+1+1+1

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

 3+1+1+1+1

 2+1+1+1+1

 7+1

 6

 4+1

 8+1

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

 +1

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

**Lesson 3: Meta(data) in Excel 15:23**

**Exercise 1: What can go wrong with data in Excel (skipped)**

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 biomass, biomass 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 15:34 15:39 15:43**

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: 10+1+1+1,+1 1+1+1

5+1+1++1

9+1+1,10+1

,7

6+1+1

·    Row 2: 7+1+1+1+1+1+1+1

10+1

5,+1

·    Column C:8+1+1, 5+1+1, 8+1

·    Column E:6 , 8+1+1, 5,6,

·    Column L:4+1+1+1, 5+1

DONE:

 +1+1+1+1

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

**Exercise 3: Outsmarted by Excel 15:45 15:49**

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? No, excel alters text to fit its own formatting assumptions noNo.NoNo

·         Can you force the above formatting? Yes.YesYesYes

·         Do you know which year these dates represent? No, NO.No!!!noNo.no

DONE: +1+1+1+1+1

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

**Exercise 4: Data tables Quiz: 15:56**

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

·         Do’s and don’ts help in programmatic re-use:TTTTTTTTTTTTTT

·         Avoiding multiple tabs improves interoperability:TTTTTTTTTTTTTTT

·         Having accompanying README file with a table description is not FAIR:FFFFFFFFFFFF

·         No ‘spaces’ in columns headers improve readability:TTTTTTTTTT/F(machine readability, yes, human readability, no)T

·         2022-07-15 is ISO date format:TTTTTTTTTTTTTT

·         20220715 date format is better for excel than 2022-07-15:FFFFFFFFFFFFF

·         “No data” is better than leaving cell “blank” for missing data:FT(in some cases)TTTFFFFFF

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

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

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

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.

-There is a lot of valuable resources in your lessons. Is there a place where all these resources are compiled?

-Will the slides be available afterwards?

-Are there any set of tools you would recommend to use together to streamline FAIR data in our experiments/work? (e.g. R together with github, etc)  - Also using electronic lab notebooks / Jupyter  (one thing I have noticed is that is difficult to automate running Jupyter notebooks)

 -Recommendations for data sharing of large datasets (text, images)

-Examples of reproducible data pipelines for example code.

- Any suggestions about ELNs? Benchling

-Can you share templates of metadata data collection?

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

-------

**Feedback:**

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

How good were the lessons:

-4444441544444445

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:4334445454333341

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:+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1

•       My knowledge has not changed much:

4.     How was the pace of the lesson:

•       Too fast:

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

•       Too slow:+1+1

5. What could be improved:

-The structure of quizes/exercises in the later lessons perhaps?

-The feedback perhaps shouldn't be semi-open like this...You could use mentimeter?

- 5 minutes wasn't enough for the breakout rooms+1+1+1

-Is Zoom the best platform?

- Would be great to have similar sessions aimed at psychologists and other academics (although I appreciate that was not the focus of todays sessions)+1

6. What did you like:

-The mention of a lot of resources that were new to me.+1

-The exercises in between+1

-Delivery from both of you+1+1+1

-Examples of metadata in repository etc

-the many examples for FAIR data and exercises to apply what we learned +1

- very clear sessions,. Great to have practical components and break out groups

Specialist subject advice

-Starting from the beginning, showing errors in data management (metadata) and showing examples of doing it correctly.