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 Biomedical Practice**

**Day 4 - Friday 31 March 2023**

**List of attendees**

- Robin Shaw, Beatson

-Caroline wise, jax

-Peter Thomason, Beatson

-Jenn Stauffer

-Rocky Onda, JAX

-Jayson Felty, JAX

-Gianluca Giusti

-Camille Liedtka, JAX

-Abigail Miller, JAX

-Cat Witmeyer, JAX

-

-

-

-

**Lesson 13: Templates for consistency**

<https://github.com/carpentries-incubator/fair-bio-practice/blob/gh-pages/files/readme_template_BioRDM.txt>

**Exercise 1: Your template**

Write your own template for either

- a measurement (for example PCR on robot)

- experiment (for example gene levels in response to stress)

In Excel or Document (txt/word) or Benchling (if you fancy using it!)

•Provide some example inputs inside your template

• remember about controlled vocabularies, PID and ontologies

(<https://fairsharing.org/>)

Folder to drop templates:

<https://uoe-my.sharepoint.com/personal/tzielins_ed_ac_uk/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Ftzielins%5Fed%5Fac%5Fuk%2FDocuments%2Ffair%2D2023%2D03%2D28&ga=1>

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

WE ARE BACK 15:05 10:05

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

**Lesson 11: Jupyter notebooks for reusable data analysis**

**Exercise 1: Basics of Jupyter Notebooks**

Open this Jupyter Notebook: <http://mango.bio.ed.ac.uk/jupyter>

We will first show you how to duplicate a notebook, save it and run code.

1.     Select the notebook titled 'student\_notebook\_light\_conditions.ipynb' and click 'Duplicate'.

2.    Confirm with Duplicate when asked if you are certain that you want to duplicate the notebook.

3.    A copy of the notebook has appeared with the suffix '-Copy' and a number. Explore the anatomy of the notebook (<https://github.com/carpentries-incubator/fair-bio-practice/blob/gh-pages/fig/10-02-jupyter_anatomy.png>)

4.    Change the title of the notebook from -copy number to your initials e.g. “student\_notebook\_light\_conditions\_IB”

5.    Save the notebook: Click on the disk symbol in the toolbar

6.    Run the notebook: Select the top cell of the notebook with the title (this is likely pre-selected already and will show with a light-blue bar to its left), and click “Run” in the tool bar.

7.    Click two times. What can you see?

-

-

-

-

8.    We want to run ALL the code: In the top tool bar click Cell > Run All. What can you see?

-

-

-

-

 DONE:

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

**Exercise 2: How to add and remove content**

You have duplicated a notebook and saved it under your own name. Now we will add text, remove cells and change code.

1.     Change the author name of the document to your name: Double click on the cell containing the author name and change the name.

2.     Press Run again.

3.     Add a new cell: Let us add details about the “light\_results.txt” file that is loaded. The “+” in the tool bar creates new cells below the currently selected cell. Thus select the cell above the code and click “+” in the toolbar.

4.     Ensure the type of the cell is Markdown and enter a description of subsequent analysis e.g.: “Loading of results following short- and long-day light exposure on arabidopsis, followed by visualisation of differences in chlorophyll/biomass etc... content between genotypes on short-days and long-days.”

5.     Press Run again.

6.     To remove a cell, select the cell you have just created and click on the scissors icon in the toolbar. (This can be undone under Edit > Undo Delete Cells)

7.     Change colours of your graph: Where the code of the graph reads the comment “# change colour of groups” you can replace the HEX codes, # followed by 6-symbol code, with names of colours (e.g. blue, green…) or other HEX codes if you are familiar with them.

8.     Save graph under new name: Add your initials to the file name under which the image is saved. Press Run. Your image should be visible in overall file hierarchy.

DONE:

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

**Exercise 3: Add another analysis step**

We have shown you how to manipulate text and code in Jupyter notebooks. You should be able to add data visualisation (a graph) and stats for long-day light condition including annotations yourself.

1.     Add additional cells including

a.     Titles

b.     Edited code to depict graph from long-days (saved under different name)

c.     Figure legend

d.     Statistical testing of difference between genotypes on long-days (remember to assign a different variable throughout e.g. LD.aov)

e.     Interpretation of results of statistical testing

DONE:

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

**Exercise 4: Sharing of your Jupyter Notebook**

You have now generated your own analysis and interpretation on top of your collaborators results and want to share this with your colleagues.

1.     Download your Notebook (ensure all code has been run) as .html and .pdf

2.    View the documents and think about why it is important to run all code before download (try Cell > All Output > Clear and download your Notebook, compare the outputs)

Questions:

What is the difference between running all code and clearing all run code?

Why is it important to run all code before download?

DONE:

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

**Exercise For Advanced**

FInd a way to plot in sensible way: both conditions LD and SD on the same graph, for the 3 genotypes and two output variables (biomas, starch).

If you created advanced plots type your name bellow:

 # plot both LD and SD using facet\_wrap

ggplot(df, mapping = aes(x = genotype, y = biomas, fill = genotype)) + # x-axis shows genotype, y-axis shows biomas

    geom\_boxplot(alpha=0.3) +

    labs(title = "Biomas per Genotype",

        x = "Genotype", # Title of x-axis

        y = "Biomas (g)") + # Title of y-axis

   # scale\_fill\_manual(values=c("#999999", "#E69F00", "#56B4E9")) + # change colour of groups

    scale\_fill\_manual(values=c("blue", "green", "red")) +

    theme\_bw() +

    theme(legend.position="none") +

    facet\_wrap(~ light\_condition)

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

**Exercise 5: Accessibility of Jupyter Notebooks**

On a scale from -2 to 2, how do you feel about the following statements (R is interchangeable with Python),

where -2 (strongly disagree), 0 no opinion to +2 strongly agree:

Type a number between -2 to 2 behind the statement:

·         making graphs for a subset of data is easier in R than in Excel: +12+`12+2+2

·         it is easier to filter for data in R than in Excel:22+2+1+2

·         it is easier to generate a series of plots with similar layout in R than Excel:2+22+2+2

·         it is easier to do large scale data processing in R than in Excel:2+22+2+2

·         using notebooks does not require any programming knowledge:-1-2-2-1-1-1

·         notebooks give you a better overview of your data analysis than Excel:2+2+2+22+2+2

·         notebooks links laboratory style records with data analysis:2+2+2+22

·         Jupyter is free, whilst a Microsoft Office (+Excel) suite costs $149.99, this alone is an incentive to use Jupyter:2+2+22+2+2

·         you need to learn R to do any data processing in notebooks:-2+111-1

·         notebooks assures reproducible computing:1+21111

·         wrong inputs or not captures parameters are main reasons for not reproducible analysis:1111

 bac

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

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

**Lesson 15: It's all about planning**

**Exercise 1: Action plan challenge**

Where would you say the following actions belong in the Research Data Life Cycle? How do they help in achieving FAIR principles? (type the corresponding number of the research data Life Cycle after each corresponding action)

**Research Data Life Cycle**

1. Creating data
2. Processing data
3. Analysing data
4. Preserving data
5. Sharing data
6. Reusing data

**Actions:**

·         Clarify usage rights:555,6565

·         Give credit through citations:55,6566

·         Use open source software:1232,3,352,35

·         Attach PID to your data:15

·         Attach descriptive metadata:121,21,2

·         Produce standard metadata:111,22

·         Backup your data:44444

·         Create figures and plots in python/R:232,323

·         Organize your files in folders:1-41,52,3,56

·         Select data repository:54,54,51,4,56

·         Add open licence:55,65

·         Link publications, data and methods:4565

·         Create a template for assay description:1,4,61

·         Use institutional repositories:4,54,55

·         Use controlled vocabularies:11,2,3,5,6

·         Convert numerical data to csv:55,65,66

·         Track versions of files:2-441,2,36

·         Performing statistical analysis:3333

·         Deposit datasets to Zenodo/Dryad:4,5,6454,54

·         Record experiment details in Electronic Lab Notebook:1,4,511,2

·         Use github for your code:4,5453,556

·         Ask someone to revise your project structure:5

·         Reformat and clean data tables:555,6

·         Use a Minimal Information Standard:1

·         Use PID in data description:155

·         Download a dataset:3,556

·         Link to UniProt or GenBank records:5

DONE:

Is there an actions above which is not clear for you or you do not know how to perform it?

 List them:

 -

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

**Exercise 2: Challenge**

**(20 minute exercise)**

Working in groups, think of your last paper (or project). Pretend that you have a joined project that combines the outputs of **at least two your papers/projects**.

You can look at the example of DMP and reusable paragraphs:

<https://www.wiki.ed.ac.uk/display/RDMS/Short+paragraphs+that+you+might+find+useful+when+preparing+your+DMP>

Our list of suggested Data Repositories can be found here: <https://www.wiki.ed.ac.uk/display/RDMS/Suggested+data+repositories>

For finding standards and repositories:

<https://fairsharing.org/>

For ontologies: <http://www.obofoundry.org/>, <https://bioportal.bioontology.org/>

Write a short DMP for this **joined project**.

**Drop the DMP document at:** <https://uoe-my.sharepoint.com/personal/tzielins_ed_ac_uk/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Ftzielins%5Fed%5Fac%5Fuk%2FDocuments%2Ffair%2D2023%2D03%2D28&ga=1>

HINT: You can drop a document there and start to collaborate on it online doing simultanous edits!

Your DMP should contain the following three sections:

**1. What data will you acquire during the project:** Please describe the type of data you will generate (for example ‘flow cytometry data’) as well as file formats and data volume. These data will be stored under (include the meta data as well).

**2. How will you store the data:**Please describe how you will store and organize your data, what metadata will you capture in what form. Explain how you will document the data during the duration of the project

**3. How will you share the data:** Please describe the strategies for data sharing, licensing and access information.

**Remember: it is a joined project**

**Exercise 2, part 2: check out each others DMPs**

(10 minute exercise)

Now, take a look at the **other group's DMP** and make comments/suggestions on how to improve it (at the end of the DMP)

**Green Room <-> Blue Room**

**Red Room <-> Yellow Room**

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

**Exercise 3: Quiz**

Answer the following questions with true or false (T or F):

1. The best time to do data management is at the end of a project, when you've collected all the data you're managing.

2. Data management plans (DMPs) detail what will happen to data before collection begins.

3. The best storage method for data is multiple backups to USBs.

4. There is a single best way to manage, organise, and share data.

5. For grant applications, DMPs should mention data preservation, longevity, sharing, discover, and reuse.

6. Your metadata should be standardised and descriptive.

7. Taking the time to plan out what's needed in metadata and your DMP will save you time in the long run and make your data more FAIR.

8. DMP online is a tool which constructs DMPs for researchers.

9. Data addressed in a DMP can be freely shared regardless of confidentiality.

10. Data can be given creative commons licenses to dictate how others can and cannot use it.

DONE:

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

**Your journey to be FAIRproductive**

**Exercise 1**

Read through the following activities / practices, type next to each

-1 if you (your group) do not perform it

? if you are not completely sure what it stands for

0 if you only learnt abou it at this workshop

+1 if you (your group) adheres / practices it

•                     include license with datasets:

•                     include license with code / scripts:

•                     use git as version control:

•                     store code in github:

•                     create DOI for datasets / code:

•                     add date availability section to a manuscript:

•                     use minimal information standards:

•                     use ontology terms:

•                     use generic data repository:

•                     use domain specific data repository:

•                     have description templates for various techniques in the lab:

•                     store data in a shared, network drive:

•                     have an automatic backup solution for files:

•                     follow a file naming conventions:

•                     create standard project folder structure:

•                     use Electronic Lab Notebooks:

•                     create figures and plots in python/R:

•                     select data repository:

•                     know non-restrictive licenses:

•                     create readme for each dataset:

•                     use institutional repositories:

•                     use controlled vocabularies:

•                     have ORCID

•                     have dedicated folder / database for protocols / SOP

•                     have a way to reference different versions of a protocol

•                     convert numerical data to csv:

•                     follow conventions for tidy data tables:

•                     use jupyter notebooks or R-markdown:

•                     use metadata format / standards:

•                     use PID from repositories (eg UniProt, GenBank) in data description:

•                     use database for bio-samples / strains etc:

•                     use pipelines for data analysis:

•                     can access all group data from your own PC:

•                     discuss with team how particular data type should be described and stored:

•                     use tools / resources you organization offers for data management:

•                     use support you organization offers for data management:

•

DONE:

**Exercise 2**

Type below the things you are going to change in your work habits or actions you are going to take after this course:

 -Be more systematic about project structure and data organisation (and teach this to others)+1+1

 - will try to learn more about coding part of data management

 -Be sure to reiterate the value of FAIR when starting initial project planning - this includes leveraging DMPs and identifying repos, etc+1

 -be more thoughtful about file, folder structures

 -Use best practices when naming metadata and data headers

 Think ahead about how we can organise and share more data+1+10000+1

 Use trackable analyses that can be shared and reused+1

 -Integrate Jupyter notebooks

 increase script documentation and traceability

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

**WHERE are all of our materials**

Our course website is:

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

It contains the text for self learnings, the excercises and the links.

There is a github repository that contains the lessons text as well as the powerpoint slides (HINT in instructors folder), but we challenge you to find the repository on your own :)

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

**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:**

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

How good was this course:

 -45444

  444

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

Was it worth your time:

-4544

 4544

On the scale 0 - 5 (zero only for my worst enemy, 5 highly recommended)

How likely are you to recommend this course to others:

-3 - depends on the person.+13

3 - it covers a lot of ground and some bits might not appeal to all+1

4444

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

•       My knowledge has not changed much:

2.      How was the pace of the lesson:

•       Too fast:

•       About right:+1+1+1+1+1+1

•       Too slow:

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

- got very confused about the templates section, so more guidance here would be good+1

-5  min intro for selfteaching or course link for coding

4. What could be improved:

-Some modules were much more suited to wet lab scientists (e.g. ELNs), some were much better for computational scientists. I think breaking the course into a wet version and a dry version would be more beneficial for both groups, or making certain modules optional depending on your needs+1+1+1+1+1

-agree with above statement, but it was very good presentatins for the both sides of data creation and management

-Further leveraging show tell do model. There were concepts i was unfamiliar with, and i would have benefitted from seeing a finished version before replicating them on my end. For example - templates, DMP. It might be useful to also have a case study to use for those with less field experience. +1+1+1+1

-I think going over pre-made examples of templates and DMPs and finding the issues would be more helpful than going over the ones we just made+1

-Should have reviewed template examples and better explained their components before the exercise

5. What did you like:

 Opportunity to see FAIR in action, meet other people+1+1

 REALLY helpful discussions with other participants+1+1+1+1

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

**Feedback For The Whole Course:**

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

How good was this course:

 455

On the scale 0 - 5 (zero useless, 5 useful)

How useful was this course:

 555

On the scale 0 - 5 (zero only for my worst enemy, 5 highly recommended)

How likely are you to recommend this course:

 444

What other topics would you add:

 It already covers a lot of ground

What topics can we shorten:

 Maybe mke the first day less totally theoretical

 A couple of people have suggested holding the course one afternoon a week for 4 weeks, for example. To give time to assimilate some of the knowledge