# FAIR for busy biologists

## Day 2

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

### List of attendees

-

-

-

--

-

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

### Your data type

Type +1 bellow the datatypes/domain which are the most relevant to your research.

1. genomics, seq data:

2. microscopy:

3. Synthetic biology, constructs

4. Proteomics

5. Metabolomics

6.  Code

7. Other (type which topic is relevant to your research):

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

### Lesson 9: Files organization

#### Exercise 1: Naming and sorting (5min + 4 explanation)

Have a look at the example files from a project, similar to the one from the previous metadata episode.

For example,

* LD\_phyA\_off\_t04\_2020-08-12.norm.xlsx

is a file that contains normalized data (norm), from experiment in long day (LD) for genotype

phyA, with media off sucrose (off).

All the files have been sorted by name and demonstrate consequences of different naming strategies.

For your information, to encode experimental details the following conventions were taken

* phyB/phyA are sample genotypes
* sXX is the sample number
* LD/SD are different photoperiodic conditions (long or short day)
* on/off are different media (on sucrose, off sucrose)
* measurement date
* other details are timepoint and raw or normalized data

       2020-07-14\_s12\_phyB\_on\_SD\_t04.raw.xlsx     (1)

       2020-07-14\_s1\_phyA\_on\_LD\_t05.raw.xlsx       (2)

       2020-07-14\_s2\_phyB\_on\_SD\_t11.raw.xlsx       (3)

       2020-08-12\_s03\_phyA\_on\_LD\_t03.raw.xlsx     (4)

       2020-08-12\_s12\_phyB\_on\_LD\_t01.raw.xlsx     (5)

       2020-08-13\_s01\_phyB\_on\_SD\_t02.raw.xlsx     (6)

       2020-7-12\_s2\_phyB\_on\_SD\_t01.raw.xlsx         (7)

       AUG-13\_phyB\_on\_LD\_s1\_t11.raw.xlsx            (8)

       JUL-31\_phyB\_on\_LD\_s1\_t03.raw.xlsx             (9)

       LD\_phyA\_off\_t04\_2020-08-12.norm.xlsx         (10)

       LD\_phyA\_on\_t04\_2020-07-14.norm.xlsx          (11)

       LD\_phyB\_off\_t04\_2020-08-12.norm.xlsx         (12)

       LD\_phyB\_on\_t04\_2020-07-14.norm.xlsx          (13)

       SD\_phyB\_off\_t04\_2020-08-13.norm.xlsx          (14)

       SD\_phyB\_on\_t04\_2020-07-12.norm.xlsx          (15)

       SD\_phya\_off\_t04\_2020-08-13.norm.xlsx           (16)

       SD\_phya\_ons\_t04\_2020-07-12.norm.xlsx          (17)

       ld\_phyA\_ons\_t04\_2020-08-12.norm.xlsx           (18)

**1 & 3 room:**

Focus on the data with date first:

       2020-07-14\_s12\_phyB\_on\_SD\_t04.raw.xlsx     (1)

       2020-07-14\_s1\_phyA\_on\_LD\_t05.raw.xlsx       (2)

       2020-07-14\_s2\_phyB\_on\_SD\_t11.raw.xlsx       (3)

       2020-08-12\_s03\_phyA\_on\_LD\_t03.raw.xlsx     (4)

       2020-08-12\_s12\_phyB\_on\_LD\_t01.raw.xlsx     (5)

       2020-08-13\_s01\_phyB\_on\_SD\_t02.raw.xlsx     (6)

       2020-7-12\_s2\_phyB\_on\_SD\_t01.raw.xlsx         (7)

       AUG-13\_phyB\_on\_LD\_s1\_t11.raw.xlsx            (8)

       JUL-31\_phyB\_on\_LD\_s1\_t03.raw.xlsx             (9)

**Questions:**

1. What are the problems with having the date first?
2. How do different date formats behave once sorted (eg 1,2 vs 8,9)?
3. Do you see what happens when you mix conventions?
4. Can you tell the importance of a leading 0 (zeros)?

**2 & 4 room:**

Focus on the other half of the files:

       LD\_phyA\_off\_t04\_2020-08-12.norm.xlsx         (10)

       LD\_phyA\_on\_t04\_2020-07-14.norm.xlsx          (11)

       LD\_phyB\_off\_t04\_2020-08-12.norm.xlsx         (12)

       LD\_phyB\_on\_t11\_2020-07-14.norm.xlsx          (13)

       SD\_phyB\_off\_t4\_2020-08-13.norm.xlsx          (14)

       SD\_phyB\_on\_t04\_2020-07-12.norm.xlsx          (15)

       SD\_phya\_off\_t04\_2020-08-13.norm.xlsx           (16)

       SD\_phya\_ons\_t04\_2020-07-12.norm.xlsx          (17)

       ld\_phyA\_ons\_t04\_2020-08-12.norm.xlsx           (18)

**Questions:**

1. Is it equally easy to find all data from LD conditions as ON media?
2. Can you spot the problem when using different cases (upper/lower) eg 15, 16, 17, 18?
3. Do you see benefits of keeping consistent lengths of the naming conventions (10-12 vs 16-17)?
4. Can you tell the importance of a leading 0 (zeros) (dated sample 13-14)?

DONE:

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

#### Exercise 2: A good name (4min)

Select which file options adhere the best to the presented recommendations:

1.

a) analysis-20210906.xlsx

b) rna-levels-by-site.v002.xlsx

c) analysis of rna levels from 5Aug2021.xlsx

2.

a) 20210906-birds-count-EDI.csv

b) birds.csv

c) birds-count&diversity EDI 2021-09-06.csv

3.

a) 2020-7-12\_s2\_phyB\_+\_SD\_t01.raw.xlsx

b) ld\_phyA\_on\_s02-t01\_2020-07-12.norm.xlsx

c) ld\_phya\_ons\_02-01\_2020-07-12.norm.xlsx

DONE:

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

#### Exercise 3: Folders vs Files (5min)

Have a look at these two different organization strategies:

(1) |-- Project

|-- |-- arab\_LD\_phyA\_off\_t04\_2020-08-12.metab.xlsx

(2) |-- Project

|-- |-- arabidopsis

|-- |-- |-- long\_day

|-- |-- |-- |-- phyA

|-- |-- |-- |-- |-- off\_sucrose\_2020-08-12

|-- |-- |-- |-- |-- |-- t04.metab.xlsx

Can you think of scenarios in which one is better suited than the other?

**Hint:**think of other files that could be present as well.

DONE:

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

#### Exercise 4: Typical folder organizations

(5 min break out, 7 min explanation)

Have a look at the four different folder structures A-D.

<https://github.com/carpentries-incubator/fair-bio-practice/blob/gh-pages/fig/07-file_organisation.png>

The first two” A) B) are recommended for computing, the other two: C) D) are for more wet/biological projects.

* Which one is the most similar to your project structure

A    B)         C)        D)

**Blue & Yellow room:**

When/why would you use A) and when/why B)

A)

B)

**Green & Red room:**

When/why would you use C) and when/why D)

C)

D)

DONE:

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

#### Exercise 5. FAIR files: (3 min)

Choose 3 main benefits of a good strategy for folder organisation and naming conventions

\* Makes data more findable

\* Aids in making data more reproducible - projects can be copied easily

\* Raw data can be reanalysed multiple times

\* Naming conventions can be read automatically

\* Easy to understand content by name, less misunderstandings

\* Easier to find and share data with others

\* Easy inspection of the project progress (present files)

\* Fewer meetings required when sharing data

\* Time saving

DONE:

### Lesson 5: Jupyter notebooks for reusable data analysis

#### Exercise 1: Basics of Jupyter Notebooks

Open this Jupyter server: <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/biomas etc... content between genotypes on short-days and long-days.”

5.     Press Run again.

6. Experiment with formatting, check the existing cells how they use (# \* - to add formatting)

7.     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)

8.     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.

9.     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?

Why share notebooks in both ipynb and html

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:

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

#### 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:
* it is easier to filter for data in R than in Excel:
* it is easier to generate a series of plots with similar layout in R than Excel:
* it is easier to do large scale data processing in R than in Excel:
* using notebooks does not require any programming knowledge:
* notebooks give you a better overview of your data analysis than Excel:
* notebooks links laboratory style records with data analysis:
* Jupyter is free, whilst a Microsoft Office (+Excel) suite costs $149.99, this alone is an incentive to use Jupyter:
* you need to learn R to do any data processing in notebooks:
* notebooks assures reproducible computing:
* wrong inputs or not captures parameters are main reasons for not reproducible analysis:

DONE:

**Lesson 6: Public repositories**

**Exercise 1a: Public general record description**

<https://doi.org/10.5281/zenodo.5045374>

We have discussed which elements of the record make it FAIR.

Now, skim through the data set description (HINT there is also a README), try to judge the following, and indicate your evaluation using marks from 0 to 5 (5 best) as to whether:

• It is clear what the content of the data set is:

• It is clear why the data could be used (i.e., what for):

• It is well described:

• How confident will you be to work with this data set:

• How easy it is to access the data set content:

• Your team datasets are equally well described (or better):

DONE:

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

**Exercise 1b: Dataset discovery**

Try to find:

- data sets related to neuromuscular junction in Zenodo

Judge the following, indicating your assessment using marks from 0 to 5 (5 best)

• how easy it is to find similar or interesting data sets:

• It is clear what the content of the other data sets are:

• It is clear why the data could be used (ie what for):

• They are well described:

DONE:

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

**Exercise 2: Domain specific repositories.**

Select one of the following repositories based on your expertise/interests:

Have a look at mRNAseq accession 'E-MTAB-7933' in [ArrayExpress]

(<https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-7933/>)

* What makes it better than Zenodo:
* What domain specific features can you see:
* Searching:

Have a look at microscopy 'project-1101' in [IDR]

(<https://idr.openmicroscopy.org/webclient/?show=project-1101>)

* What makes it better than Zenodo:

-

-

-

* What domain specific features can you see:

-

-

-

* Searching:

-

-

-

Have a look at the synthethic part record 'SubtilinReceiver\_spaRK\_separated' within the 'bsu' collection in [SynBioHub](<https://synbiohub.org/public/bsu/SubtilinReceiver_spaRK_separated/1>)

* What makes it better than Zenodo:

-

-

-

* What domain specific features can you see:

-

-

-

* Searching:

-

-

-

Have a look at the proteomics record 'PXD013039' in [PRIDE]

(<https://www.ebi.ac.uk/pride/archive/projects/PXD013039>)

* What makes it better than Zenodo:
* What domain specific features can you see:
* Searching:

Have a look at the metabolomics record 'MTBLS2289' in [Metabolights](<https://www.ebi.ac.uk/metabolights/MTBLS2289/descriptors>)

* What makes it better than Zenodo:
* What domain specific features can you see:
* Searching:

DONE:

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

**Exercise 3: Finding a repository (3 min +3)**

Our own curated repository list:

<https://www.wiki.ed.ac.uk/display/RDMS/Suggested+data+repositories>

Using Fairsharing (<https://fairsharing.org/>) find a repo for flow cytometry data and type the name below:

-

-

-

once done, search for repository for genomics data

-

-

-

**------------------------------------------------------------------------------------------------------------------**

**Exercise 5: Wrap up discussion**

Discuss the following questions:

* Why is choosing a domain specific repositories over zenodo more FAIR?
* How can selecting a repository for your data as soon as you do an experiment (or even before!) can benefit your research and help your data become FAIR?
* What’s your favourite research data repository? Why?

DONE:

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

**Your journey to be FAIRproductive**

**Exercise 1**

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

-1 if you do not perform it

? if you are not completely sure what it stands for

0 if you only learnt about it at this workshop

+1 if you adheres / practices it

• include license with datasets:

• include license with code / scripts:

• use git as version control:

• create DOI for datasets / code:

• add date availability section to a manuscript:

• use minimal information standards:

• 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 convention:

• 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 controlled vocabularies:

• have ORCID

• have dedicated folder / database for protocols / SOP

• have a way to reference different versions of a protocol

• follow conventions for tidy data tables:

• use jupyter notebooks or R-markdown:

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

• use database for bio-samples / strains etc:

• can access all group data from your own PC:

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

• use support your 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:

-

-

-

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

## WHERE are all of our materials

Our big course website (FAIR in bio practice) is:

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

It covers more materials and often in more depth than this shorter workshop.

Fair for busy website is:

And the slides and exercises are in ‘instructors’ folder on git:

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

Q&A:

Do you have any questions about the topics discussed 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:

-

-

-

-