# FAIR in (Biological) Practice

Link to course

## Day 2

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

#### Important notice:

Before you begin today, please take some time to sign up for the following two accounts:

1. Benchling (the ELN we will use for one of today's lessons): <https://benchling.com/signup?pubref=pubref_zQlS6DPe>.
2. Protocols.io (the protocol repository with PID, which we will also be using for today's lessons): <https://www.protocols.io/create>

(do not worry if you cannot do it, right now... there will be some time to do this at the start of lesson 8)

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

### List of attendees

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

### Your data type

Type your name 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 6: Ontologies

Finding ontologies:

<https://bioportal.bioontology.org/>

List of recommended ontologies:

<http://www.obofoundry.org/>

Exploring ontologies

<https://bioportal.bioontology.org/ontologies/ZFA/?p=classes&conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FZFA_0001109#details>

<https://bioportal.bioontology.org/ontologies/CL?p=classes&conceptid=CL:0000129#details>

<https://bioportal.bioontology.org/ontologies/SO?p=classes&conceptid=http://purl.obolibrary.org/obo/SO_0000167#details>

Exercise

1. The prefix CL stands for:
2. Class ontology:
3. Cell ontology:
4. Cell line ontology
5. The recommended ontology for chemical compounds is:
6. cheminf
7. chmo
8. chebi
9. Which terms captures both Alzheimer’s and Huntington's diseases
10. DOID\_680
11. DOID\_1289
12. DOID\_0060090

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

### Lesson 7: 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:

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

### Lesson 8: Laboratory Records

**Which of these most accurately describes your record keeping experience?**

(type +1 next to the statement that best describes your situation)

- I have used hard copy lab notebooks before for research data records keeping.

- I have used electronic lab notebooks before for research data records keeping.

- I have used Benchling before.

- I have used an online protocol database before.

- I have used Protocols.io before.

- I have not had any research record keeping experience in the past.

- What in the world are you talking about? Is this Philosophy 101 or am I in the wrong classroom?

DONE:

Before you begin with lesson 8, please take some time to sign up for the following two accounts (if you haven't already done so):

1. Benchling (the ELN we will use for one of today's lessons): <https://benchling.com/signup?pubref=pubref_zQlS6DPe>.
2. Protocols.io (the protocol repository with PID, which we will also be using for today's lessons): <https://www.protocols.io/create>

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

#### Exercise 1:

**Differences between analog and digital record keeping**

Compare the electronic version of the tea protocol:

<https://www.protocols.io/view/how-to-make-a-cup-of-tea-buhknt4w>

with the paper one from the photo:

<https://github.com/carpentries-incubator/fair-bio-practice/blob/gh-pages/fig/06-handwritten-tea-protocol.jpg>

What are advantages and disadvantages of traditional analog records vs digital records? Try to find at least a handful of advantages and disadvantages for each. With all of these, which system do you think is most advantageous?

Green & Red room

**Advantages of traditional analog records**

**Advantages of digital records**

Blue & Yellow room

**Disadvantages of traditional analog records**

**Disadvantages of digital records**

DONE:

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

#### Exercise 2:

**Re-using a published lab entry**

1.     Open Benchling ([https://benchling.com](https://benchling.com/)) and log in.

2.     First within your own workspace click the big ‘+’ (Create Project) right next to Projects in your Benchling workspace

3.     Call the project ‘Breakfast’, and add an appropriate description, click ‘Create project’

4.     Click here <https://benchling.com/s/etr-SY8fi7L8ZIDSMCLCf92o> to access the public lab entry ‘Eggs Florentine in Portobello Mushrooms’.

5.     Select the clock symbol on the right-hand side underneath Share: Now you can see the history of the entry and changes that have previously been made to the document with a timestamp. If someone had tried to ‘manipulate’ data, you would be able to see this here. You also see the owner of the document.

6.     Click ‘Clone from version’.

7.     Select the ‘Breakfast’ folder to clone it to.

DONE:

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

#### Exercise 3:

**Adapting a protocol to your needs**

1.     You have now accessed a digital record and want to reuse it to make your own breakfast. To show how reusable digital records are we will first navigate through the cloned file you made in your project.

2.     Navigate to your Project ‘Breakfast’, you can tell you are in your Project, if your initials show in a red circle next to entries in the side bar. You should see the lab entry ‘Eggs Florentine in Portobello Mushrooms’, and the top bar above the title and toolbar should read ‘Tea’, ‘Portobello Mushrooms and Spinach’, ‘Poached Egg and Hollandaise Sauce’, ‘Add Protocol’, ‘Notes’, and ‘Metadata’.

3.     Click through those tabs and you will see that in your notes you have your lab entry describing how breakfast was made with embedded graphics and a shopping list and current prices. The other three tabs describe the protocols that were used, and you can add additional protocols with the ‘add protocol’ tab. We want you to adapt the ‘Tea’ protocol to suit your ingredients and methods.

4.     Once you have made appropriate changes in the Tea protocol, you should consider changing the order in which the breakfast and tea are made.

5.     Once you have made all suggested change have a look at the history of the record (clock button), you can see the changes you have recently made, and you can see it still relates to the original document. It tells you what record it has been cloned from and when.

6.     Click the link to the original record. As you can see digital record keeping allows provenance, crediting the original author, but also allowing you to keep track of your sources.

7.     Navigate back to your lab entry in your project (your initials are a sign that you are in the right place).

DONE:

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

#### Exercise 4:

**Sharing your record**

1.     Click the info icon on the right-hand side underneath the clock symbol you used previously and select ‘Export entry’

2.     Your export is now running, you will receive an email when the export is complete

3.     Click the link in the email to download your protocol as a .zip

4.     Unzip the file and in your own time, print the protocol if you want to use the recipe in the kitchen, or share it with friends.

5.     You can share .pdf versions or click Share and generate a Share link of your lab entry. This makes your record interoperable as many users across many platforms across the world can access your entry if you make it public and share it on for example social media. If there is no IT access present, you always have the option to print the .pdf copy.

DONE:

Benchling demo:

<https://benchling.com/s/etr-D59zgfqSvefvhPmIxAne?m=slm-FOH9KYfv9n6eUZPJWbkt>

Plasmid example (plasmids private accessible for UoE BioRDM only)

<https://benchling.com/s/etr-Gje9DwS83aijjPlkmaH9?m=slm-0EcbtGdiD4I9xpcfOByr>

**Our Benchling tutorial:**

<https://www.wiki.ed.ac.uk/display/RDMS/Benchling+%28quick%29+tutorial>

**Further resources and tutorials from Benchling:**

* Main help page, with access to several tutorials: <https://help.benchling.com/en/>
* More molecular biology features: <https://help.benchling.com/en/collections/69523-molecular-biology>
* Benchling training kit for academics: <https://help.benchling.com/en/collections/1608962-benchling-training-kit-for-academics>

**Materials on ELNs:**

<https://www.wiki.ed.ac.uk/pages/viewpage.action?pageId=463750271>

DONE:

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

#### Exercise/challenge 5:

**Adapt a public protocol and retain its provenance**

Protocol link: <https://dx.doi.org/10.17504/protocols.io.buhknt4w>

*Fork the protocol, preserving the original for crediting*

1. Open the link to the above protocol, as you can see, we have assigned it its own DOI

2. First click on Metrics: Because we are FAIR, this shows you how many views over time this protocol has had, how many exports, how many individual steps it involves and how many times it has been forked.

3. Now click on the downwards arrow next to the title

4. Select 'Copy/Fork' and click 'make a fork'

5. Select the Folder you want the protocol to be forked to and click 'continue'

6. Your fork of "How to make a cup of tea" is ready now, click 'edit new fork'

7. On the right-hand tool bar, the clock icon, shows you the history of the protocol (as before in Benchling). Currently you should see no history as you have not made changes.

*Edit the forked protocol*

1. Go to 'Materials' in the top tool-bar: add or edit materials according to your preferences, e.g., change full-fat milk to oat-milk, or add honey, lemon etc

2. Go to 'Steps' in the top tool-bar: edit the protocol according to your preferences

3. You can edit the 'Description' and 'Guidelines & Warnings' if you would like to

4. As soon as you change anything, the timestamp and where in the protocol this change was made appears in the history.

5. Click 'View', you will now see the reader view of your protocol. It clearly states underneath the title 'Forked from How to make a cup of tea' and the original protocol is linked. This allows clear identification of your source.

6. Click 'Edit'

*Optional: Export the forked protocol*

1. Click 'More' in the top tool-bar, select 'Export' > 'PDF' > 'To your computer' and click export (leave selections blank)

DONE:

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

#### Exercise 6:

**Do you use an ELNs? Which one? What features do you like?**

* HAVE YOU TRIED AN ELN and you gave up?:
* Anyone switched from paper entirely to ELNs?:
* Anyone only paper?:

Which ELN do you use:

-

-

-

DONE:

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

#### Quiz:

Which of the following statement are true (T) / false (F)?

* Good record keeping ensures transparency and reproducibility.
* There are no advantages to using analog record keeping when compared to digital record keeping.
* Digital records help people view a protocol simultaneously.
* Digitally kept records can be quickly and easily edited.
* On balance, digital record keeping is more advantageous than analog record keeping.
* Digital records are easier to search (for and within) than analog records.

DONE:

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

### Lesson 9: Files organisation

#### Exercise 1: Naming and sorting

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)

**Blue & Green 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)?

**Red & Yellow 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\_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)

**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 1-3)?

DONE:

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

#### Exercise 2: A good name

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

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

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:

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:

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

#### 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: