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 (Biological) Practice <https://edcarp.github.io/2022-02-15_ed-dash_fair-bio-practice/>

**Day 2**

(Day 3 notebook <https://pad.carpentries.org/2022-02-17-ed-dash-fair>)

**List of attendees**

1. Kadi Vaher
2. Adelaide Young
3. Amelia Edmondson-Stait
4. Flávia Fonseca Pezzini
5. Wu Huang
6. T Zhou
7. Nneka Nnadi
8. Caity Ellis
9. Marta Amador
10. Cigdem Selli
11. Sumy Vnnilathil Baby

**Your data type**

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

1. genomics, seq data:

.fastq (RNA-seq)+1+1+1(single cell)+1

.fast5 (ONT)

metagenomics (Kadi)

Plant Genomics

2. microscopy:

.tiff (DIC images) / .czi (confocal microscopy) +1+1

3. Synthetic biology, constructs

.gb (plasmid constructs)

4. Proteomics

 +1+1

5. Metabolomics

6.  Code

R+1+1+1+1+1, perl, bash+1+1

 Python+1+1

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

 large population cohorts eg. UK Biobank (Amelia)

 birth cohorts (Kadi)

 neuroimaging (Kadi)+1

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

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

**Lesson 6: Being precise**

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

**Registries**

•species e.g. NCBI taxonomy

<https://www.ncbi.nlm.nih.gov/Taxonomy>

•chemicals e.g. ChEBI

<https://www.ebi.ac.uk/chebi>

•proteins e.g. UniProt

<https://www.uniprot.org/>

•genes e.g. GenBank

<https://www.ncbi.nlm.nih.gov/genbank/>

•metabolic reactions, enzymes e.g KEGG

<https://www.genome.jp/kegg/>

**Lesson 6: Ontologies**

Finding ontologies:

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

List of recommended ontologies:

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

Excercise 3.

1.       The prefix CL stands for:

a)       Class ontology:

b)      Cell ontology: +1+1+1+1+1+1

c)       Cell line ontology

2.       The recommended ontology for chemical compounds is:

a)       cheminf

b)      chmo

c)       chebi +1+1+1+1

3.       Which terms captures both Azheimer’s and Huntington's diseases

a)       DOID\_680

b)      DOID\_1289+1+1+1+1

c)       DOID\_0060090

DONE:

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

**Lesson 7: Meta(data) in ExcelExercise 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?

 colour pattern has no explanation of what they mean+1+1+1

 also bold font has no explanation

-what do the acronyms/abbreviations mean+1

-Some Missing data+1

-values and units together

- missing units+1

-different options for errors/missing values

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

 - Sample in first table seems to correspond with Cell in second table, but that also has a sample column

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

 Different unit(mg/g) and some entries missing their units

 units together with the values in the cells

- 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?

-colours would not be saved in csv/txt formats

-the spreadsheet is not machine readable - looks like two tables in one, units together with values in cells

**Answers:**

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

+1 especially when people don't do stats in R or other similar programming language

DONE:

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

**Exercise 2: Spotting problems 14:04**

Which of these problems (repeated below and numbered) are apparent in 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>

in the following places:

(add the corresponding numbers to the end of these bullet points)

·    Row 5: 5, 7, 1, 9

·    Row 2: 7, 11

·    Column C: 8, 5,1,

·    Column E: 6. 8, 5

·    Column L: 5, 4

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

DONE:

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

**Exercise 3: Outsmarted by Excel 14:14**

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+1+1+1+1+1

·       Can you force the above formatting? Yes +1+1+1 '

·       Do you know which year these dates represent?Not clear (2020 maybe?)

DONE:

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

**Exercise 4: Data tables and FAIR**

How do the described practices for representing data in tables (Excel, .csv or .tsv) help in achieving FAIR? Which aspects of FAIR do they help with?

DONE: 14:25

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

 WE ARE BACK 14:45

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

**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. +1+1+1+1+1+1

- I have used electronic lab notebooks before for research data records keeping.+1(github/gitlab)+1+1(labarchives)+1

- I have used Benchling before.+1

- I have used an online protocol database before.

- I have used Protocols.io before. +1

- I have not had any research record keeping experience in the past. +1 because I've not done research

- 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?

**Advantages of traditional analog records**

 -don't need internet to make notes (or a laptop/tablet if in a wetbench lab)+1

 -

 -not limited to the formatting of the digital notebooks (e.g. can draw, annotations etc)

**Advantages of digital records**

 - versions normally automatically saved/timestamped

 -easily shared with others

handwriting not ligiable

easier to copy methods from other digital media

- can easily search/file by experiment+1

- Can easily attach files/put links to folders

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

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

**Exercise 3: 3:10**

**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:+1+1+1+1+1+1+1+1+1

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

**Exercise 4: 15:17**

**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: +1+1+1+1+1+1

**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:+1+1+1+1+1+1+1+1

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

**Exercise 6: 3:34**

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

HAVE YOU TRIED AN ELN and you gave up?: we tried R space as a group but majority did not want to use it and prefered using their own record keeping mainly because it was difficult and time consuming to export already written protocols into the new platform

Labarchive/ one note/ slack

Anyone switched from paper entirely to ELNs?:+1 I enjoy accesing my protocols from anywhere at anytime

Anyone only paper?+1: +1+1 (Currently - I used lab archives when my previous institute switched to it)

DONE:+!

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

**Quiz:**

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

·       Good record keeping ensures transparency and reproducibility.TTTTTTT

·       There are no advantages to using analog record keeping when compared to digital record keeping.FFFFFFF

·       Digital records help people view a protocol simultaneously.T TTTTTT

·       Digitally kept records can be quickly and easily edited.TTTTTTT

·       On balance, digital record keeping is more advantageous than analog record keeping.TTTTTTT

·       Digital records are easier to search (for and within) than analog records.TTTTTTT

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

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

 Back 4:50

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

**Lesson 9: Files organisation**

Exercise 1: Naming and sorting 3:54

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)

**Room 1:**

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?

 not necessarily reflect the structure of experimental design

1. How do different date formats behave once sorted (eg 1,2 vs 8,9)?

1. Do you see what happens when you mix conventions?

1. Can you tell the importance of a leading 0 (zeros)?

**Room 2:**

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?

Overall it's ok to tell the difference between LD/SD except lowercase for last entry (18).

1. Can you spot the problem when using different cases (upper/lower) eg 15, 16, 17, 18?

Not as easy to search? eg. using regular expressions

1. Do you see benefits of keeping consistent lengths of the naming conventions (10-12 vs 16-17)?

there is a mixture of 'on' and 'ons'

 if there's a difference between on and ons you cannot easily tell apart

1. Can you tell the importance of a leading 0 (zeros) (dated sample 1-3, above)?

 Keeps numbers in order

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+1+1+1+1+1+1

c) analysis of rna levels from 5Aug2021.xlsx

2.

a) 20210906-birds-count-EDI.csv+1+1+1+1+1+1

b) birds.csv

c) birds-count&diversity EDI 2021-09-06.csv+1

3.

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

b) ld\_phyA\_ons\_s02-t01\_2020-07-12.norm.xlsx+1+1+1+1+1+1+1

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

DONE:+1+1+1+1

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

**Exercise 3: Folders vs Files 4:12**

Have a look at these two different organization strategies:

(1) |-- Project

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

if you have too many files in the same folder?

If you only have this type of data with the same variables in this folder you could use this structure otherwise it would become confusing+1

sharing files with other people

(2) |-- Project

|-- |-- arabidopsis

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

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

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

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

if you are woorking with multiple organisms for instance, different times

you can include each protocol in this scenario

i only use this if the study is way too big

easily missed up if you move files + sometimes need to change the directory in R too often which is confusing!

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 4:17**

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)

**Room1:**

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

A)easy to put the script folder in the path for analysis

 easy to upload repository and leave out "data" folder if can't share this due to sensitive data

can easily avoid not touching the raw data folder (so that don't accidentally edit/delete)

B)

 maybe different students working on different aspects of the larger project

 directly feed the data into the script

 overall I think this is mostly preference.

**Room2:**

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

C)for comparing between individuals

we prefer this layout!

D) for comparing between different time points

DONE:

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

**Exercise 5. Discussion:**

How can a strategy for folder organisation and naming convention help in achieving FAIR data? (Add +1 if you agree with any of the sentences written by your colleagues):

If it is easy to understand the data will be easier to find and understand so it can be reused+1

 If you plan it, it is easy to document it to share+1+1

 Our lab share folder has been messy it is because no one has ever written a readme file

 If there is a logical structure it will be more accessible (eg. not spread in random places)

 What about the data structure #A, then R has to alawys keep changing the working directory? -> no. you can just refer to the directories in your script. the 'here' package is very good for that if you are working with R projects

 Very hard to keep different versions of scripts that are slightly different eg. one with more detailed notes but a bit messy

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

**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.

How good is onenote for lab notebook? some of my colleagues use it

 Would you say that it is worth having a person dedicated to data management? Or at what point/project size do oyu think having the people generating the data also managing it becomes unfeasible? - some lab managers do that (big labs)

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

**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:+!+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

•       Too slow:

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

 longer breaks

 The thing that get changed by excel - and how to stop it /reverse it

4. What could be improved:

5. What did you like: interactive format, relevant exercises, rigth pace, room for discussion+1+1 (much better discussions today - possibly the slightly bigger groups today is why?)+1+1+1