Files organisation Episode Notebook

Part of FAIR in (bio) practice, <https://carpentries-incubator.github.io/fair-bio-practice>

Type your name and institution:

-

-

-

**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 light 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 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)

·         What are the problems with having the date first?  
·         How do different date formats behave once sorted (eg 1,2 vs 8,9)?  
·         Do you see what happens when you mix conventions?

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

Green room:

Focus on the other half of 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:

·         Is it equally easy to find all data from LD conditions as ON media?  
·         Can you spot the problem when using different cases (upper/lower) eg 15, 16, 17, 18?  
·         Do you see benefits of keeping consistent lengths of the naming conventions (10-12 vs 16-17)?

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

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

**Folders vs Files**

Have a look as 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 other? Hint: think of other files that could be present as well.

-

-

-

Typical folder organizations

Have a look at the four different folder structures.

1. <https://github.com/carpentries-incubator/fair-bio-practice/blob/gh-pages/fig/10-a_computing_org_birds.png>
2. <https://github.com/carpentries-incubator/fair-bio-practice/blob/gh-pages/fig/10-b_computing_org_birds_by%20experiment.png>
3. <https://github.com/carpentries-incubator/fair-bio-practice/blob/gh-pages/fig/10-c_bio_org_by_pig.png>
4. <https://github.com/carpentries-incubator/fair-bio-practice/blob/gh-pages/fig/10-d_bio_org_by_treatment.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 room:

* When/why would you use a) and when/why b)

Green room:

* When/why would you use c) and when/why d)

----------- removed -------------

**Organization for computing**

Take a look at the folder structure recommended by the Good enough practices in scientific computing paper.

Why do you think this layout is recommended and suited for a computing project?

.

|-- CITATION

|-- README

|-- LICENSE

|-- requirements.txt

|

|-- data

| |-- birds\_count\_table.csv

|

|-- doc

| |-- notebook.md

| |-- manuscript.md

| |-- changelog.txt

|

|-- results

| |-- summarized\_results.csv

|

|-- src

| |-- sightings\_analysis.py

| |-- runall.py

|

----------- removed -----------------

**Feedback**

**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.      Thinking of your knowledge of the lesson topic and its presentation,

which one of the statements best characterize your experience (type +1

next to the statement)

•       I am a novice, and I found the course useful/informative:

•       I am a novice, but I think the course should be improved:

•       I have experience in the presented area, but I found the course

useful/informative:

•       I have experience in the presented area, and I think the course could

be improved:

3. How was the pace of the lesson:

•       Too fast:

•       About right:

•       Too slow:

4. If the lesson had to be 5 minutes shorter, what would you remove:

-

-

-

5. If the lesson could be 5 minutes longer, what would you add or spend

more time on:

-

-

-