# FAIR for busy biologists

## Day 2

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

### 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 9: Files organization

#### 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)

**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

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 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:

-

-

-

-