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 Biomedical Practice**

**Day 3 - Thursday 30 March 2023**

**List of attendees**

-Camille Liedtka, JAX

-Jenn Stauffer, JAX

Peter Thomason, Beatson

-Caroline Wise

 Robin Shaw, Beatson

Jayson Felty, JAX

-Rocky Onda, JAX

-Gianluca Giusti

-

-Abigail Miller, JAX

-

-Cat Witmeyer, JAX

-

-

-

-

-

**Your coding experience:**

Please, enter +1 if:

0: Not much coding experience: +1+1+1+1

1. You use jupyter notebooks: +1+1

2. You sometimes write scripts (bash etc): +1+1

3. You sometimes code in R:+1 +1+1 +1

4. You sometimes code in Python:+1 +1+1

5. You code in other language (type which): +1

First we will complete an exercise from yesterday.

**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 +1  +1 +1       B)  +1 +1 +1     C)       D)+1+1

Room 1:

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

A) A for a perhaps purely software related project - the work is organized into accessible/high-level categories which allows for the work to be accessible to more external collaborators more quickly - however it doesn't distinguish by type of experiment.

B) This is also good for lab work (from my perspective). It separates data by experiment type, which can be very usefl

 This structure is supportive of data collection and analysis over time (such as a field season)folder by figure

Rooms 2 & 3:

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

C)Structured like a paper so helpful when preparing a manuscript

Easy to do automated analysis when all results are in a single folder

Easier to grab all data for individual specimen

D)Possibly earlier on in study when organizing by chronology is helpful and don't need to structure things for a paper yet

DONE: 1

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

**Lesson 11: Jupyter notebooks for reusable data analysis**

**Exercise 1: Basics of Jupyter Notebooks**

Open this Jupyter Notebook: <http://mango.bio.ed.ac.uk/jupyter>

Do you have a Jupyter Notebook open yet?+1+1+1+1++11+1++11+1

is the text large enough? Would you like it to be larger?

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

Do you have your own notebook open, the one with your initials?+1 ++11+1+1+1+1+1

Are you seeing blinking things at upper right?Yes

No+1+1+1+1+1

Firefox OK for me+1

I was but disconnected from VPN and it worked.

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.     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)

7.     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.

8.     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?

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:

You can try to create the same plots in python (but we cannot guarantee all the packages are installed)

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

**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 12: Version Control**

**Exercise 1: Problems with a change**

Which of these issues can you relate to? (Type +1 next to each statement if you relate to it)

·         I have fifteen versions of this file - which one do I use?++11+1+1+1+1++1

·         I can’t remake this figure from last year.+1

·         I slightly modified my code in one place, everything stopped working.+1+1+

·         I have several copies of the same directory because I'm worried about breaking something.+1

·         Somebody added erroneous records in a shared file with samples, cannot find who and why+1+1+1

·         You remember seeing a data file but cannot find it anymore: is it deleted? Moved away?+1+1+1+1+1+1+

·         I tried multiple analysis and I don't remember which one I chose to generate my output data+1+1

·         I have to merge changes to our manuscript from 10 different emails with collaborators+1+1+1+1

·         I accidentally deleted a part of my work +1+1

·         I came to an old project and forgot where I left/saved it+1+1+1

·         I reorganized my data files, but broke the analysis pipeline and don't know how to fix it+1

·         I have trouble to find the source of a mistake in an experiment

·         My directory is polluted with a lot of unused/temporary/old folders because I'm afraid of losing something important +1+1

·         I made a lot of changes to my paper but want to bring back one paragraph+1

DONE: +1+1+1+1+1+1+1+1

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

**Exercise 2: Manual Versioning**

Which of these issues does manual versioning help to tackle?

(Type +1 next to each statement if manual versioning helps to tackle)

·         I have fifteen versions of this file and I don't know which one to use ++11+1+1+1+1+1

·         I can't remake this figure from last year +1+1+1

·         I slightly modified my code in one place, everything stopped working++11+1+1+1+1+1+1

·         I have several copies of the same directory because I'm worried about breaking something+1+1+1+1

·         Somebody added erroneous records in a shared file with samples, I cannot find who and why+1+1 if they add to the change log

·         I tried multiple analysis and I don't remember which one I chose to generate my output data+1+1+1+1+1

·         I have to merge changes to our manuscript from 10 different emails from collaborators ·         I made a lot of changes to my paper but want to bring back one paragraph+1+11+1

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

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

**Exercise 3: Changelog in action**

Have a look at one of the example Github repositories and how they track changes\*:

Green 1 and Blue 2:

·         data from E.R. Ballou et al. 2020

·         <https://github.com/ewallace/pseudonuclease_evolution_2020/commits/master>

Give examples of:

·         what makes them a good changelog

 -The changes are granular & provide extensive detail

 Author and date are tracked

·         what could be improved

 -Adding explanation for why the changes were made

 -No commit comments

 -No branching

 -Sometimes TOO much information makes it difficult to distinguish what changes were made & why

Which are the most difficult features to replicate with manual version control?:

 Harder to undo changes/backtrack

Red, Yellow 3:

·         data from I. Boehm et al. 2020

·         <https://github.com/BioRDM/nmj-pig/commits/main>

Give examples of:

·         what makes them a good changelog

 describes what was changed

 know who made the change and when

·         what could be improved

Which are the most difficult features to replicate with manual version control?

 DONE:

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

**Exercise 4:**

Five reasons to use a version control system in research, give +1 to the two most important for you

·         Tell the story: The history of your commit messages will describe your project progress. +1+1+1

·         Travel back in time: a version control system makes it easy to compare different time points of your project smoothly. If you want to compare the stage of your project a year ago from now, it only takes one command-line of code +1+1+1

·         Experiment with changes: if you want to make changes in a script you can first make a “snapshot” of the project status before experimenting with changes.+1+1+1

·         Backup your work: by being able to link your local repository (folder) to a distant online host (GitHub), a version control system backs up your precious work instantly+1+1+1

·         Collaborate easily on projects: having a web-hosted synchronised version of your project will encourage collaboration with other researchers. Think about a colleague of yours being able to add a script to make a figure for your publication for instance. The repository will track their contribution and evidence their participation in the project +1+1+1+1+1

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

 BACK 4:10, 11:10

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

**Learning GIT:**

 •Software Carpentry git workshop <https://swcarpentry.github.io/git-novice/>

•Edinburgh Carpentries runs courses: <https://edcarp.github.io/>

•<https://ourcodingclub.github.io/tutorials/git/>

•Crash course on youtube: <https://youtu.be/SWYqp7iY_Tc>

•Learn git branching: <https://learngitbranching.js.org/>

•Learn git-game: <https://github.com/git-game/git-game>

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

**Semantic versioning quiz**

Type +1 after the correct answer:

1. Which of the library version is the latest?

·         a) 0.12.4

·         b) 1.1.0

·         c) 1.12.3-alpha+1+1 ++11+1+1+1+1

·         d) 1.12.2

2. You use library fair.2.3.1 to automatically upload your data after analysis.  You found out that there is available new version of the library fair.4.3.1. Which situation is the most probable?

·         a) fair.4.3.1 uploads data twice as fast as fair.2.3.1

·         b) you need to reserve time to test the new version and adapt your analysis pipeline to work with it +1+1+1+1+1+1+1

·         c) both versions offer the same features and bug fixes as they have the same minor/path version 3.1

3. You developed code that helps to plot detrended and normalized data, the last release has version 1.2.1. You added Z-score to the available normalization methods and fixed the spelling errors in the user messages. Which version number should you give to the new release:

·         a) 2.0.0

·         b) 1.3.0+1+1+1+1+1+1+1+1

·         c) 1.3.2

·         d) 1.2.2

·         e) 2.2.1

4. Your group releases data about studied cancer patients following semantic  versioning. The last data sets are released as 1.11, you added data from a new patient, what version should you use:

·         a) 1.12+1+1+1+1+1+1+1

·         b) 2.11

·         c) 2.0+1

 5. You were asked to better anonymize the above cancer data (v1.11) in future releases. Instead of data and time of patient visit you release only the year, the hospital names have been encoded and patient age has been     obscured by random noise. What version should you use:

·         a) 1.12

·         b) 2.1

·         c) 2.0+1+1+1+1+1+1+1+1

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

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

**Lesson 13: Templates for consistency**

<https://github.com/carpentries-incubator/fair-bio-practice/blob/gh-pages/files/readme_template_BioRDM.txt>

**Exercise 1: Your template**

Write your own template for either

- a measurement (PCR on robot)

- experiment (gene levels in response to stress)

In Excel or Document (txt/word) or Benchling (if you fancy using it!)

\* provide some example data

Folder to drop templates:

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

**Lesson 14: Public repositories**

**Exercise 1: Public general record**

Have a look at the following record for data set in Zenodo repository:

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

1. What elements make it FAIR?

FINDABLE (persistent identifiers, easy to find data and metadata):

ACCESSIBLE (The (meta)data retrievable by their identifier using a standard web protocols):

INTEROPERABLE (The format of the data should be open and interpretable for various tools):

REUSABLE (data should be well-described so that they can be replicated and/or combined in different settings, reuse states with a clear licence):

2. Skim through the data set description (HINT there is also a README), try to judge using marks from 0 to 5 (5 best) if

·         It is clear what the content of the data set is: 35444554

·         It is clear why (what for) the data could be used:44445554

·         It is well described:4533454-+3

·         How confident will you be to work with this data set:33354353

·         How easy it is to access the data set content:45555555

3. Give +1 to the statement that the best describes your latest public dataset or a current project folder:

* It does not have readme like description: +1+1(abstract)+1
* It description has only one/two paragraph(s), I have not thought     of adding so many details: +1
* My data set description is similar in details:
* My data /project description is much richer or formalised:

DONE:

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

**Exercise 2: Dataset discovery**

Try to find either:

- similar data sets in Zenodo

- data sets of interest for you

Judge using marks from 0 to 5 (5 best)

·         how easy is to find similar or interesting dataset:545454

·         It is clear what the content of the data set is:54344

·         It is clear why (what for) the data could be used:4334

·         They are well described:43343

DONE:

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

**Exercise 3: 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:

 Can download large sequencing data

 Specically for sequence data

·         What domain specific features can you see:

 Navigation and menus

·         Searching:

 Faceted search is domain specific

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

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

·         What makes it better than Zenodo:

-specifically for images, has image browser, tools to manipulate images

Robust metadata

-really nice image preview!

-

·         What domain specific features can you see:

-has specific fields for types of images, like cell lines, phenotypes

-

-

·         Searching:

-not great? -- I see a sarscov2 experiment but when I search for "sars" nothing comes up

-

-

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:

 easier access to information as species or disease etc.

 more studies recorded

·         What domain specific features can you see:

·         Searching:

DONE:

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

**Exercise 4: Finding a repository**

Firstly, check the publisher's / funder' recommended list of repositories, some of which can be found below:

- [BioMed Central / Springer Nature](<https://www.springernature.com/gp/authors/research-data-policy/recommended-repositories>)

- [eLife]([https://submit.elifesciences.org/html/elife\_author\_instructions.html#policies](https://submit.elifesciences.org/html/elife_author_instructions.html" \l "policies))

- [Elsevier](<https://www.elsevier.com/about/policies/research-data>; <https://www.elsevier.com/authors/tools-and-resources/research-data/data-base-linking>)

- [EMBO Press]([https://www.embopress.org/page/journal/14602075/authorguide#datadeposition](https://www.embopress.org/page/journal/14602075/authorguide" \l "datadeposition))

- [F1000 Research](<https://f1000research.com/for-authors/data-guidelines>)

- [GIGAscience - OUP](<https://academic.oup.com/gigascience/pages/instructions_to_authors>; <https://academic.oup.com/gigascience/pages/editorial_policies_and_reporting_standards#Reporting%20Standards>)

- [PLoS](<https://journals.plos.org/plosbiology/s/recommended-repositories>)

- [Scientific Data - Nature](<https://www.nature.com/sdata/policies/repositories>)

- [Taylor and Francis](<https://authorservices.taylorandfrancis.com/data-sharing-policies/repositories/>)

- [BBSRC](<https://bbsrc.ukri.org/research/resources/>)

- [NERC](<https://nerc.ukri.org/research/sites/environmental-data-service-eds/policy/>)

- [Royal Society](<https://royalsociety.org/journals/ethics-policies/data-sharing-mining/>)

- [Wellcome Open Research](<https://wellcomeopenresearch.org/for-authors/data-guidelines>)

Secondly, check [Fairsharing recommendations](<https://fairsharing.org/recommendations/?q=>)

- alternatively, check the [Registry of research data repositories - re3data](<https://www.re3data.org/>)

BioRDM's curated list of repos: <https://www.wiki.ed.ac.uk/display/RDMS/Suggested+data+repositories>

a) Find and type a repo for genomics data:

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

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

<https://ngdc.cncb.ac.cn/gsa/>

Cancer Genomics Hub <https://www.re3data.org/repository/r3d100011174>

b) Your favourite/chosen data type and a recommended repo for it:

<https://www.informatics.jax.org/> :)+1+1

c)   List repositories you have used (either to download or to share your data): Figshare, Zenodo

 MGI, FlyBase, GenBank, Ensembl, ENCODE, MGI

DONE:+1+1+1

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

**Exercise 5: Wrap up discussion**

Discuss the following questions:

·         Why is choosing a domain specific repositories over zenodo more FAIR?

 It allows the work to be more findable to the right community+1

better minimal metadata for better accessibility

·         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?

You can structure your metadata, readmes, etc according to the standards of the repository which will help encourage FAIRness +1+1

 data becomes available earlier

·         What’s your favourite research data repository? Why?

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:+1+1+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++11+1

•       Too slow:

3. From 0 to 5 how useful was the jupyter demo:

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

- perhaps another short exercise with repositories

5. What could be improved:

-During breaks, instructors could pop into breakout rooms to discuss lesson planning+1

-

-

6. What did you like:

 Livia did a great job getting us back on track!  :)+1+1+1

 Playing around in the repositories was helpful

 Y'all did a good job pivoting despite technical issues+1