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).

Users are expected to follow our code of conduct: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html

---

Useful links

- Zoom room for this workshop:-https://zoom.us/j/98816109991 the password was given in the emails you received prior to the workshop.
- Workshop webpage: https://edcarp.github.io/2020-04-29-sfc-online/
- Dataset:
- Data Skills Workforce Development https://www.ed.ac.uk/bayes/about-us/our-work/education/workforce-development
- Code of Conduct https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html
- Pre-workshop survey: https://carpentries.typeform.com/to/wi32rS?slug=2020-04-29-sfc-online
- Spreadstreets lesson material: https://datacarpentry.org/spreadsheet-ecology-lesson/
- OpenRefine lesson material: https://datacarpentry.org/OpenRefine-ecology-lesson/
- SQL lesson material: https://datacarpentry.org/sql-ecology-lesson/
- Post-workshop survey: https://carpentries.typeform.com/to/UgVdRQ?slug=2020-04-29-sfc-online

Attendance list

Please write your name to confirm your attendance at the workshop:

1. Martina Benkova x x
2. Gabrielle Migdalska x x
3. Pooja Jain x x
4. Chris Osborne x x
5. Julie Corrigan x x
6. Kirsten Anderson x x
7. Ian Horobin x x
8. Joe Marshall x x

9. Hari Nair xx
10. Alex Cuadrado x x
11. Serafim Alvanides x x
12. Agata Podvorska

EDINBURGH CARPENTRIES

https://edcarp.github.io/ -- sign up to the mailing list
bottom of the hompage)
get in touch with g.peru@epcc.ed.ac.uk

Timetable

Time schedule Wednesday

9:00-09:30 Intro

- 9:30-10:45 Data organisation with Spreadsheets
- 10:45-11:15 *Coffee break 1*
- 11:05-12:30 Data cleaning with Spreadsheets
- 12:30-13:30 *Lunch break*
- 13:30-15:00 OpenRefine
- 15:00-15:30 *Coffee break 2*
- 15:30-17:00 OpenRefine

Time schedule Thursday

- 9:00-10:30 SQL
- 10:30-11:00 *Coffee break 1*
- 11:00-12:30 SQL
- 12:30-13:30 *Lunch break*
- 13:30-15:00 SQL
- 15:00-15:30 *Coffee break 2*
- 15:30 -17:00 SQL

Day 1 morning

PPT

https://drive.google.com/open?id=1VA_Xdy8ZvPeL7yqxPsWD50kdRu8E-9hd

Lessons 1: Data organisation with spreadsheets

- You will need any spreadsheet program, e.g. Excel, OpenOffice, Google Sheets, LibreOffice
- Data
    - Data Messy: https://ndownloader.figshare.com/files/2252083
    - Cleaned dataset https://github.com/edcarp/2020-04-29-sfc-online/raw/gh-pages/files/Survey_cleaned.xlsx
    - Semi cleaned dataset https://github.com/datacarpentry/spreadsheet-ecology-lesson/blob/gh-pages/data/survey_sorting_exercise.xlsx?raw=true

Data sets for today: https://figshare.com/articles/Portal_Project_Teaching_Database/1314459

CSV file format

.csv format: https://www.lifewire.com/csv-file-2622708
CSV = comma-separated values format - a way to save spreadsheets in plain text files (each line is one raw and values in rows are separates with commas), e.g.:

Name,weight,height
Harry,85,190
Mia,60,165
Dennis,75,183

Formatting issues

- using text and numbers in cells, different date formatting, mixing variables in rows and collumns
- Inconsistencies not only between the two researchers/years, but also within the same onw, esp 2014!
- There were different formatting in different cells, colour used to encode data, two data variables combined in one column (species and sex)
- empty cells - missing value or 0?
- Finished but a very manual (error-prone!) process
- Also multiple tables in a single sheet
- The resulting spreadsheet will still need lots of reformatting and sorting to be any use
- issues with dates appearing as a number rather than date format also one of them was 1978?
- Text and numbers used together in same cells for measurement units
- Multiple variables combined in same column (2014 sheet Plot 4)
- Colour used to encode calibration data
- Date discrepancies

Re: metadata file template - notes:

Difficult to get enough for people for that minimal core of required metadata isI think the Dublin core is the minimal dataset that was in agreement

There are different standards depending on the domain or application

Also depending to which istitution you belong to there must be different standard

Following is a good starting point that you may consider to adopt:
https://data.research.cornell.edu/content/readme

Date formatting/ using formulae to extract Y/M/D:

This example article might provide a bit additional insights:
https://www.extendoffice.com/documents/excel/3118-excel-extract-month-and-year-from-date.html

The best way it to either use 3 separate columns or tell Excel to treat your columns as 'text' and it won't apply the automatic conversion to Date data type

So if you tell Excel to treat a column as text, then you can use a date format you are comfortable with, e.g. one of the standard ones are 2020-04-29 YYYY-MM-DD

Might be interesting: https://answers.microsoft.com/en-us/office/forum/office_2007-excel/how-do-i-use-dates-in-excel-before-1900-in-a/0a152162-3e64-e011-8dfc-68b599b31bf5

There are some upsides to using dates as data type - you can do some operations with dates, e.g. extract rows with a date before a specific date or you can add dates and do some arithmetic operations on dates Depends on your needs - just be aware of the pitfalls

Feedback for the morning session

Can you please come up with one thing you liked about the morning session on spreadsheets and one thing you did not like or that could be improved. Please add these as bullet points.

Thing I liked

- Hari - The Excel data cleaning exercise was very helpful. The tips gave a clear picture on how to go about the process.
- Yes, data tidying was very useful
- Very useful to see data validation
- I learned a lot this morning - some very useful tips
- Explanation clear. I like the contente explained is useful in practice.
- Very helpful, no complaints
- In general I thought the online Zoom format worked very well and I enjoyed the workshop. Things like practical exercises work just as well online as in a class room. I once working on one mointor, but it was a 27" montitor and everything worked fine for me.

Thing that could be improved

- Hari - some more advanced exercises could be included.
- Obviously, working from home made this an issue, but I found trying to work on one laptop screen a bit fiddly; hopefully it won't be more of an issue the rest of the way
- Still rather confused about how to tell Excel whether dates are US or UK style
- Just working from home and web cam issues -the instruction is great.
- More practice would be welcome. Maybe also with different types of tables.
- Very helpful, no complaints
- It was a wee bit frustrating at times to use the Raise Hand non-verbal badge and get no response. Also using one of your recomeneded spread sheet application was a bit confusing as some of your instructions were Excel specifc.

Day 1 afternoon

Lesson 2: Data cleaning with OpenRefine

- PPT: https://drive.google.com/open?id=1ZX5ymi97ithN1VTTRBzUbUcFRb8SpUGs
- Installation instructions: https://datacarpentry.org/OpenRefine-ecology-lesson/setup.html
- Data:
    - https://ndownloader.figshare.com/files/7823341 and
    - https://github.com/data-curation/cdcf-workshop2A/blob/master/SAFI_openrefine.csv (data is part of the Portal Project Teaching Database available on FigShare https://figshare.com/articles/Portal_Project_Teaching_Database/1314459)

Importing your data:

- when importing excel it gives you a choice of what worksheets you want to upload (all by default). I just tried.

Working with lists:

- Al
- value.split(";", "")

Clustering algorithms used:

Clustering algorithms used by OpenRefine: https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

Feedback for the afternoon session

Can you please come up with one thing you liked about the afternoon session on OpenRefine and one thing you did not like or that could be improved. Please add these as bullet points.

Thing I liked

- Good pacing overall, especially given the complexity of some tasks.
- Really good to get a sense of what Open Refine can do
- OpenRefine is very useful. Not used it very much yet I'm more confident using it after this.
- The topics covered were well chosen to get a sense of the potential of OpenRefine. It was well explained and the sharing-screenshot feature from Lucia was so useful and used properly.
- An excellent overview of OpenRefine. I liked that I could work through the exercises with Lucia. Very glad I can return and go over the exercises myself to practice what I have learned today.
- Very useful tool to learn. Really got a good sense of what OpenRefine is capable of.

Thing that could be improved

- When an exercise is set, please make sure that it stays on the screen so that we can see it for the duration. Alternatively, post the exercises in the Group chat or maybe here (at the pad)
- Final hour was quite stretching but that might just be the consequence of 6 hours in front of the screen!
- An awful lot to take in. As mentioned above the last bit was a bit of a trial. Perhaps better spread over several sessions?
- I missed more practice in some of the exercises of the last hour. The fact that we can ask questios in the chat during the explanation distracted me a bit but that's a personal thing I guess. For future workshops I'd really encourage particpants to use two screens to follow the explanations as we practice (maybe a training with different options for this issue might be included in the preparatory sesion we have the day before the course starts).
- I honestly can't suggest anything. The instructors are doing such a great job from their own homes. Thank you so much.
- It was a great session. The only thing that could be included is an initial outline of what Data wrangling entails and requires - a brief overview of the process and then putting that in the context of OpenRefine.

Day 2

Lesson 3: SQL

- Installation instructions: https://datacarpentry.org/sql-ecology-lesson/setup.html
- **Download** this data to your computer: https://doi.org/10.6084/m9.figshare.1314459. Click on **Download all** to download the zip file. Unzip it to a location that you can easily find on your computer.
- SQL lesson material: https://datacarpentry.org/sql-ecology-lesson/

Data files for this session - direct download links (avoid file compression e.g. .zip):

SQLite file:
https://ndownloader.figshare.com/files/11188550
Plots: https://ndownloader.figshare.com/files/3299474
Species: https://ndownloader.figshare.com/files/3299483
Surveys: https://ndownloader.figshare.com/files/10717177

Other – useful sources

https://www.sqlite.org/datatype3.html

https://stackoverflow.com/questions/tagged/sql

https://www.w3schools.com/sql/default.asp

https://dbschema.com

SQL (structured query language) formatting (syntax, semantics)

SELECT column FUNCTION(column) AGGREGATE_FUNCTION(column)
FROM table
JOIN table ON table.col = table.col

- WHERE (conditional statement, applies to row values) (AND/OR) (IS (NOT) NULL)
- GROUP BY column

HAVING (conditional statement, applies to group)
ORDER BY column/FUNCTION(column) (ASC/DESC)

There is a rule book, but it may look a bit scary at the moment: https://www.sqlite.org/lang_select.html

****

SELECT DISTINCT species_id from surveys;

SELECT DISTINCT year, species_id from surveys;

SELECT year, species_id, weight from surveys;

SELECT year, species_id, round(weight/1000,2) from surveys;

SELECT year, month, day, species_id, weight*1000 from surveys;

SELECT * from surveys where species_id="DM";

SELECT * from surveys where sex="F";

```
SELECT  * from surveys where year > 2000;

SELECT distinct year from surveys;

SELECT * from surveys where (year > 2000) and (species_id="DM");

SELECT * from surveys where (species_id = "DM") or (species_id="DO") or (species_id="DS") or
(species_id ="DX");

SELECT day, month, year, species_id, weight/1000 from surveys where (plot_id = 1) and (weight > 75);
```

It is also more readable and shorter as your queries grow bigger and more complescomplex

```
SELECT * from species;

SELECT * from species order by taxa, species ASC/DESC;

SELECT species_id, avg(weight), count(*)
FROM surveys
GROUP BY species_id;

SELECT year, species_id, count(*) "Count",round(avg(weight), 2) "Average weight"
FROM surveys
GROUP BY year, species_id;

select year, species_id, count(*) "Count",round(avg(weight), 2) "Average weight"from surveys group by
year, species_idorder by count(species_id) DESC;

select year, species_id, count(*) "Count",round(avg(weight), 2) "Average weight"from surveys group by
year, species_idorder by year ASC, count(species_id) DESC;
```

When using ‘aliases’ for aggregate function columns, quotation is not necessary (the syntax depends on many factors) But the code is certainly more readable and reproducible when using ‘AS’ when naming specific columns, particularly, when a query become more complex

```
select species_id, count(species_id) as occurances
from surveys
group by species_id
having count(species_id) >10
```

WHERE is practically doing the same thing as HAVING, but while WHERE is filtering actual existing column values, HAVING is to filter the results of the aggregating functions

```
select *
from surveys
where year = 2000 AND month >= 6 AND month < 10

select count(*)
from (select * from surveys where year = 2000 AND(month >= 6 and month < 10));
```

```
select surveys.year, surveys.month, surveys.day,species.genus, species.species
from surveys
join species
using(species_id);

select s1.year, s1.month, s1.day, s2.genus, s2.species
from surveys as s1
join species as s2
using(species_id);

select s1.year, s1.month, s1.day, s2.genus, s2.species
from surveys as s1
join species as s2
where s1.species_id = s2.species_id;

select plot_id, count(genus)
from surveys
join species
on surveys.species_id = species.species_id
group by plot_id, genus;

SELECT species, ROUND(AVG(weight),2)
FROM surveys
JOIN species
ON surveys.species_id = species.species_id
WHERE taxa="Rodent"
GROUP BY species;
```

From: https://www.codeproject.com/Articles/33052/Visual-Representation-of-SQL-Joins


```
SELECT species, ROUND(AVG(weight),2)
FROM surveys
JOIN species
ON surveys.species_id = species.species_id
WHERE taxa="Rodent"
GROUP BY species;
```


Feedback for day 2

Can you please come up with one thing you liked about the day 2 on SQL and one thing you did not like or that could be improved. Please add these as bullet points.

Thing I liked

- The introduction to DB Browser was very useful, even for an old SQL hack like me :-)
- The Into to Databases and DBMS was helpful. I wasn't aware of how to create a DB from CSVs in DB Browser and editing datatypes. That helped a lot.
- Very satisfying to actually write a SQL query
- It was a great overview of the possibilities with SQL.
- Good instroduction to SQL
- I agree, I found SQL to be quite a useful addition; I've seen it lurking behind MS Access but never really figured it out. Now I can!
- Great overview of SQL and good to have a chance to practice queries
- Great introduction to SQL and learnt some very handy queries. The instructors were very helpful

Thing that could be improved

- You have no choice but to stick to the commands that work in SQLite in the workshop, but it might be worth pointing out to participant that SQL can change quite a lot depending on the RDBMS you're using (e.g. as a user of Microsoft products I am used to the TOP rather than LIMIT command) as this could catch them out later if they use something other than SQLite.
- Maybe a few advanced exercises. for SQL I understand this is an intro course but I would also suggest a follow-up course with more advanced SQL lessons if there is sufficient interest in the audience. It would also help for people with smaller screens (and no second screen) if you gave the instructions on how to use multiple windows at the very beginning of the first day.
- I would have preferred to spend more time practising the more straightforward queries. I couldn't do any of the afternoon exercises and this was a bit discouraging. I felt there was more assumed knowledge than advertised - for those of us who don't have much experience with relational or mathematical data, the concepts were pretty alien.
- I think selecting a dataset with not so technical concepts about ecology, just trivial things would make the understanding easier.
- I would suggest less topics for the day with longer explanations. I was a complete beginner to SQL so I got thrown off by certain concepts that I’m sure more experienced users would have found very basic. I agree with one of the comments above that I felt that there was a little bit of assumed knowledge, and I’m not completely new to coding in general. I also agree that using a dataset about basic items (ie tables and chairs or dogs, cats, etc) would have made the process a little easier as well.
- It felt a bit rushed (how you can do it in 3 hours I don't know). The syntax is quite complicated and I found I was forgetting bits of it when we got something new. The one screen thing is a bit of a problem!
- As a complete beginner I would also have preferred more time gaining confidence in creating simple queries but it was good to see the more advanced aspects so I can explore these further on my own

In addition, if a colleague were to attend one of these virtual sessions what one piece of advice would you give them to make the best of the course?

I guess the obvious (but most important!) one is you really need either a second monitor or a reasonably big monitor to make best use of an online workshop like like this. Also a quite room where you won't be disturbed, or alternatively a headset.
To install ahead of time and familiarise a bit with the tools used in the workshop and to practice using

multiple windows (especially placing them side-by-side) if you have a smaller screen.
Get the software and data files downloaded several days in advance so you can get help if needed
Find out how to have two screens.
Like everyone else - two screens!
Two screens a must!
Two screens!