

ETC5512: Instruction to Open Data

Lecturer: Kate Saunders

Table of contents

Learning Objectives	1
Before your tutorial	1
Package Installation	2
Exercise 1: Atlas of Living Australia	2
a. Download	2
b. Data quality checks	3
c. Data collection methods	3
Learning to code	4
Additional Exercises	4

Learning Objectives

- Utilise and access open data sources
- Assess the collection methods and the quality of the data
- Write code to conduct quality checks of the data
- Also learn your way around RStudio.

Before your tutorial

Ensure you have installed both **R** and **RStudio**.

- Follow theses [installation instructions](#)

Also and work through the following [startR modules](#):

- Do the module on RStudio Basics (Module 2).
- Do the module on R Basics (Module 3).

These should take you ~ 40 minutes.

Package Installation

Run the below code to install the packages we need for today's tutorial.

```
install.packages("tidyverse")
install.packages("galah")
```

Once the packages are installed, load the packages using the `library()` function.

```
library(tidyverse)
library(galah)
```

Installing packages is like screwing in a lightbulb. Loading the package from the library is like turning the switch on and off.

You only need to screw in the lightbulb once (install the package), but you can turn the light on many times (load the library).

Exercise 1: Atlas of Living Australia

The Atlas of Living Australia is a major resource for occurrence data on animals, plants, insects, fish.

a. Download

- i. Point your browser to <https://www.ala.org.au>. Check the terms of use. Does it have a license?
- ii. Using the `galah` library, and the function `occurrences` extract the records for platypus. To download the data from this API **you will need to register with your email** first. Once you've done that change the code below so it has your email.

```
library(galah)

# add your email address below
galah_config(email = "ADD_YOUR_EMAIL_HERE",
             download_reason_id = 10,
             verbose = TRUE)

platypus <- galah_call() |>
  galah_identify("Ornithorhynchus anatinus") |>
  atlas_occurrences()
```

Take a look at the data you downloaded.

```
View(platypus)
```

iii. Let's do some data wrangling to tidy up our data.

```
platypus <- platypus |>
  rename(Longitude = decimalLongitude,
         Latitude = decimalLatitude) |>
  mutate(eventDate = as.Date(eventDate)) |>
  filter(!is.na(eventDate)) |>
  filter(!is.na(Longitude)) |>
  filter(!is.na(Latitude))
```

iv. Save our result. To save both your data and the file you are working on in the same place, in the top menu go to Session > Set Working Directory > To Source File Location. (If you are already familiar with R Projects you may like to use that instead.)

```
platypus_file_name = "platypus.csv"
write_csv(platypus, file = platypus_file_name)
```

b. Data quality checks

i. Plot the locations of sightings. Where in Australia are platypus found?

```
read_csv(platypus_file_name)

ggplot(platypus, aes(x=Longitude, y=Latitude)) +
  geom_point()
```

ii. What dates of sightings are downloaded?

```
platypus |> select(eventDate) |> summary()
```

c. Data collection methods

How is this data collected? Explain the ways that a platypus sighting would be added to the database. Also think about what might be missing from the data?

Learning to code

Tip

When you are getting started it can be useful to check your basic understanding of what each piece of code does.

Some tips:

- Use the help menu to look up what functions do, e.g. `?write_csv`
- Copy and paste code chunks in Generative AI and ask it to explain what the code is doing.

Then go back and add comments to your code to explain what the different pieces do.

Additional Exercises

- i. Try finding examples of each of the different flavours of data we discussed in class.
 - Find the license information
 - Look at the meta data.
- ii. Also be sure to complete the Open Data Institute Quizzes from Lectures to consolidate your understanding.

© Copyright Monash University