**Manual data collection**

**The scraping process**

R code: (websitename.R)

Working Directory: (website name folder)

1) Identify a website, save the URL of the homepage

2) Extract all the urls of each plant species “factsheet” or webpage into a list in r

3) Identify which data you would like to extract about the plant, then code R to open each url, download the specific elements you need and put them into a table format. At this stage, the table format is:

|  |  |  |  |
| --- | --- | --- | --- |
| study | species | trait\_name | value |
| (to identify the website) | The species as is written on the website | The trait as a heading written on the website, or summarised by me | The word used by the website to describe the trait |

4) Write the file out as a csv file titled the same as the website name

**Cleaning scraped data**

Working Directory: (website name folder)

1) Briefly check the data in excel

2) Save as AAAv1.csv

3) Begin the R analysis: AAAv1.R and open the file AAAv1.csv

4) Get a list of all the traits that are recorded by the website. For example: watering, flower colour, etc.

5) Adapt the data to the format we have chosen in our trait list. This may take many steps depending on the data format presented in the website

4) Save the file as AAAv2.csv

**Collating and species matching**

**Editing the database and summarising**