

SANBI-GBIF training workshop in Data Management and Cleaning:

Getting GBIF data in R

Vernon Visser



SANBI 
Biodiversity for Life 

Getting GBIF data using rgbif

We will look at how to obtain occurrence data from the Global Biodiversity Information Facility. As you will have learned from the previous session on GBIF, there are a number of ways of interrogating GBIF data (e.g. date, region, IUCN red list status...).

We will use an example of downloading data for a grass genus (*Pentameris*) in South Africa.



Registering

The first step will be to create your own GBIF account. Once you have registered, log in to GBIF using the username and password you chose.

The screenshot shows the GBIF (Global Biodiversity Information Facility) homepage. The top navigation bar includes links for 'Get data', 'How-to', 'Tools', 'Community', and 'About'. A prominent banner features the text 'Free and open access to biodiversity data' over a background image of hanging plant pods. Below the banner, there are four main statistics: 'Occurrence records' (1,908,372,255), 'Datasets' (64,535), 'Publishing institutions' (1,776), and 'Peer-reviewed publications' (6,573). On the right side, a registration form is displayed with fields for 'COUNTRY', 'EMAIL', 'USERNAME', and 'PASSWORD'. Below these fields are 'NEXT' and 'OR' buttons, followed by social media sign-up options: 'SIGN UP WITH GOOGLE' (Google icon), 'SIGN UP WITH FACEBOOK' (Facebook icon), and 'SIGN UP WITH GITHUB' (GitHub icon). At the bottom of the registration form, there is a link to the 'Ebbe Nielsen Challenge'.



Overview of what we will do

We will first look at how one downloads data directly from GBIF. This should help you understand the various options available on the portal.

We will then look at ways of **downloading data from GBIF in R**. We will use the `rgbif` package to do this.

First step - search

The first step will be to search for a particular taxon. Although this is possible from the GBIF home page, we are using the occurrences page because this is how you will search using filters, such as by country or publisher...

The screenshot shows a web browser window with the URL gbif.org/occurrence/search?occurrence_status=present&q=Pentameris. The page is titled "Occurrences". On the left, there is a sidebar with filters for "Occurrence status" (Present), "Licence", "Scientific name" (Pentameris), and a detailed list of subgenera: **Pentameris P.Beauv.**, **Genus**; **Pentameris E.Mey.**, **Genus**; **Pentamerista Maguire**, **Genus**; **Pentamerismus McGregor, 1949**, **Genus**; and **Pentameris scandens (H.P.Linder) Galle**, **Genus**. The main content area displays a table of search results with columns: Scientific name, Country or area, Coordinates, Month & year, Occurrence status, Basis of record, and Data source. The results include entries for *Rytidosperma racemosum*, *Lerista apoda*, *Regulus regulus*, *Sorghum leiochladum*, *Arctocephalus forsteri*, *Chroicocephalus ridibundus*, *Callitris glaucophylla*, *Lycopodium fastigiatum*, *Parus major*, *Dendrocopos major*, *Pica pica*, and *Corvus cornix*.

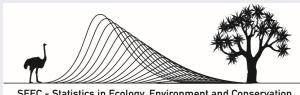
Scientific name	Country or area	Coordinates	Month & year	Occurrence status	Basis of record	Data source
<i>Rytidosperma racemosum</i> (R.Br.) Connor & ...	Australia	35.2S, 149.0E	2023 January	Present	Human observation	Nat
<i>Lerista apoda</i> Storr, 1976	Australia	17.9S, 122.3E	2023 January	Present	Human observation	All
<i>Regulus regulus</i> (Linnaeus, 1758)	Denmark	55.3N, 10.8E	2023 January	Present	Human observation	Sp
<i>Sorghum leiochladum</i> (Hack.) C.E.Hubb.	Australia	36.2S, 149.0E	2023 January	Present	Human observation	Nat
<i>Arctocephalus forsteri</i> (Lesson, 1828)	Australia	34.9S, 135.7E	2023 January	Present	Human observation	Ear
<i>Chroicocephalus ridibundus</i> (Linnaeus, 1766)	United Kingdom of Great ...	52.6N, 0.7W	2023 January	Present	Human observation	All
<i>Callitris glaucophylla</i> J.Thompson & L.A.S.J...	Australia	32.5S, 146.7E	2023 January	Present	Human observation	NS
<i>Lycopodium fastigiatum</i> R.Br.	Australia	35.5S, 148.8E	2023 January	Present	Human observation	Nat
<i>Parus major</i> Linnaeus, 1758	Russian Federation	55.9N, 38.1E	2023 January	Present	Human observation	RU
<i>Dendrocopos major</i> (Linnaeus, 1758)	Russian Federation	55.9N, 38.1E	2023 January	Present	Human observation	RU
<i>Pica pica</i> (Linnaeus, 1758)	Russian Federation	55.9N, 38.1E	2023 January	Present	Human observation	RU
<i>Corvus cornix</i> Linnaeus, 1758	Russian Federation	55.9N, 38.1E	2023 January	Present	Human observation	RU



Occurrences page

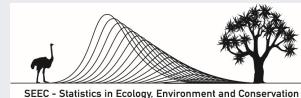
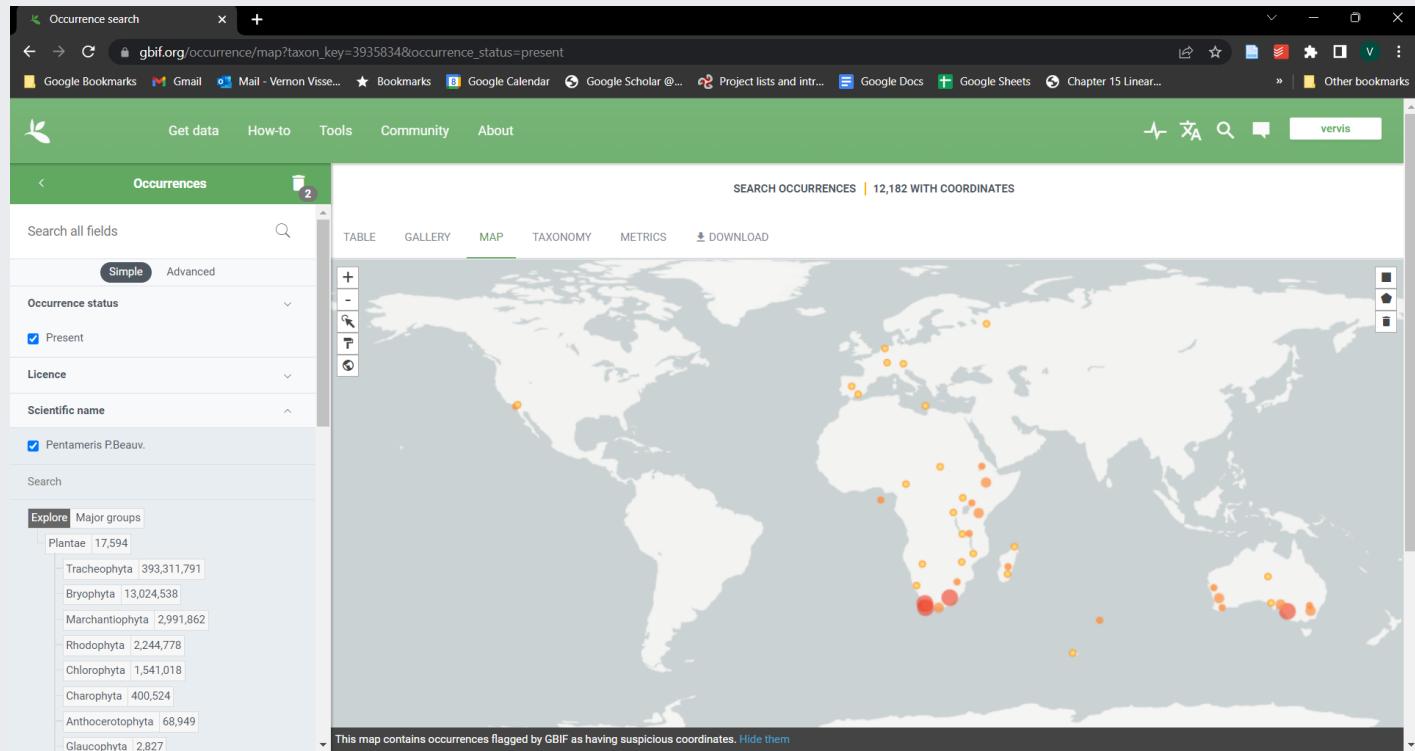
This is what the main occurrences search and filter page looks like. On the left you can see all the filter options available. Above the table of occurrences, you can see various tabs, which we will look at in the next few slides.

The screenshot shows a web browser window for the GBIF Occurrences search page. The URL is gbif.org/occurrence/search?taxon_key=3073. The page has a green header with tabs for Occurrences, Get data, How-to, Tools, Community, and About. On the far right of the header are icons for a heart, a user profile, a magnifying glass, and a message bubble, followed by a 'Login' button. Below the header is a sidebar on the left containing filter options: Occurrence status (Simple, Advanced), Licence, Scientific name (with a checked Poaceae checkbox), Basis of record, Location, Administrative areas (gadm.org), Coordinate uncertainty in metres, Year, Month, Dataset, Country or area, Continent, and Issues and flags. To the right of the sidebar is a table titled 'SEARCH OCCURRENCES | 35,307,992 RESULTS'. The table has columns for Scientific name, Country or area, Coordinates, Month & year, Basis of record, and Dataset. The first few rows show occurrences for Phyllostachys Siebold & Zucc. from Slovenia and Norway, and Brachypodium sylvaticum from Norway. At the bottom of the table are navigation links for TABLE, GALLERY, MAP, TAXONOMY, METRICS, and DOWNLOAD. The bottom of the page features a Windows taskbar with icons for File Explorer, Google Chrome, Mail, and others, along with system status indicators for battery level, signal strength, and date/time (16:02, 2022/01/05).



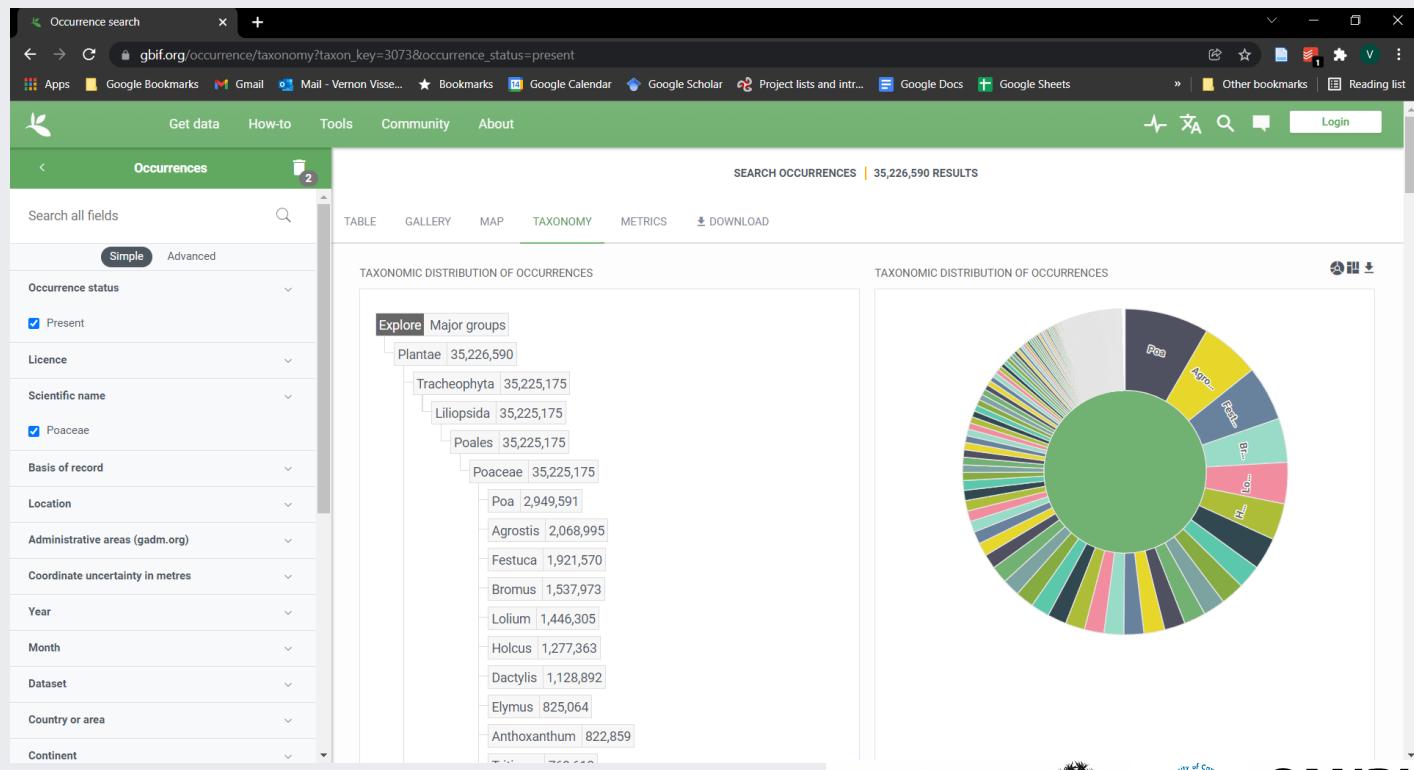
Map of occurrences

In one of the occurrences tabs you can view the distribution (map) of the particular selection of occurrences you have made.



Taxonomy of occurrences

In another of the tabs you can view the taxonomy of the particular selection of occurrences you have made. This is an important one to check to make sure you have selected the correct taxon and not used e.g., a taxon in a completely different group of animal or plants with the same name, an outdated name...



Filtering by country

For our particular case we are going to filter *Pentameris* occurrences to just those in South Africa. You do this by selecting the "Country" filter on the left-hand side and searching for and selecting "South Africa".

The screenshot shows the GBIF Occurrence search interface. On the left, there is a sidebar with filters for 'Occurrences'. Under 'Country or area', the 'South Africa' checkbox is selected. The main panel displays a table of occurrence records. The first few rows are:

Scientific name	Country or area	Coordinates	Month & year	Occurrence status	Basis of record	Date
Pentameris pallida (Thunb.) Galley & H.P.Lin...	Australia	34.9S, 138.9E	2023 January	Present	Human observation	INa
Pentameris pallida (Thunb.) Galley & H.P.Lin...	Australia	34.9S, 138.9E	2023 January	Present	Human observation	INa
Pentameris eriostoma (Nees) Steud.	South Africa	34.0S, 20.1E	2023 February	Present	Human observation	INa



SANBI
Biodiversity for Life



South African National Biodiversity Institute

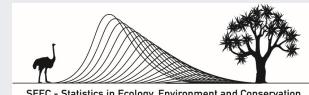
Pentameris occurrences in South Africa

You will see that the number of occurrences in the data table will drop by quite a bit.

Screenshot of the GBIF Occurrence search interface showing Pentameris occurrences in South Africa.

The search results table displays 9,027 results for Pentameris species in South Africa. The columns include Scientific name, Country or area, Coordinates, Month & year, Occurrence status, Basis of record, and Data reference. The results show various species like *Pentameris eriostoma*, *Pentameris thunbergii*, *Pentameris pallida*, etc., with coordinates ranging from 33.5S to 34.9S and 18.3E to 20.1E, mostly from 2022 and 2023.

Scientific name	Country or area	Coordinates	Month & year	Occurrence status	Basis of record	Data ref.
<i>Pentameris eriostoma</i> (Nees) Steud.	South Africa	34.0S, 20.1E	2023 February	Present	Human observation	iNat
<i>Pentameris thunbergii</i> P.Beauv.	South Africa	34.0S, 21.2E	2022 May	Present	Human observation	iNat
<i>Pentameris pallida</i> (Thunb.) Galley & H.P.Lin...	South Africa	34.7S, 19.9E	2022 September	Present	Human observation	iNat
<i>Pentameris pallida</i> (Thunb.) Galley & H.P.Lin...	South Africa	34.1S, 18.9E	2022 September	Present	Human observation	iNat
<i>Pentameris pallida</i> (Thunb.) Galley & H.P.Lin...	South Africa	34.4S, 19.5E	2022 September	Present	Human observation	iNat
<i>Pentameris curvifolia</i> (Schrad.) Nees	South Africa	34.4S, 21.4E	2022 September	Present	Human observation	iNat
<i>Pentameris curvifolia</i> (Schrad.) Nees	South Africa	34.1S, 18.4E	2022 October	Present	Human observation	iNat
<i>Pentameris airoides</i> (Nees) Steud.	South Africa	33.5S, 18.3E	2022 October	Present	Human observation	iNat
<i>Pentameris curvifolia</i> (Schrad.) Nees	South Africa	34.0S, 25.6E	2022 October	Present	Human observation	iNat
<i>Pentameris heptameris</i> (Nees) Steud.	South Africa	34.0S, 25.7E	2022 October	Present	Human observation	iNat
<i>Pentameris eriostoma</i> (Nees) Steud.	South Africa	33.7S, 23.1E	2022 October	Present	Human observation	iNat
<i>Pentameris densifolia</i> (Nees) Steud.	South Africa	33.9S, 19.2E	2022 December	Present	Human observation	iNat



Downloading - log in

At this stage, if you have not already logged in, you will need to first log in to be able to download.

The screenshot shows a web browser window for the GBIF Occurrence search at gbif.org/occurrence/download?country=ZA&taxon_key=3073&occurrence_status=present. The interface includes a sidebar with filters for Occurrence status (Present), Licence (Poaceae), Basis of record, Location, Administrative areas (gadm.org), Coordinate uncertainty in metres, Year, Month, Dataset, and Country or area (South Africa). The main search results show 190,131 results. A prominent 'SEARCH OCCURRENCES' button is visible. A 'DOWNLOAD' tab is selected, showing options for TABLE, GALLERY, MAP, TAXONOMY, METRICS, and DOWNLOAD. A large modal dialog box is overlaid on the page, containing a 'LOGIN' section with fields for 'USERNAME OR EMAIL' and 'PASSWORD', a 'Forgot your password?' link, a 'SIGN IN' button, and social media login options: 'CONTINUE WITH GOOGLE' (Google icon), 'CONTINUE WITH FACEBOOK' (Facebook icon), 'CONTINUE WITH GITHUB' (GitHub icon), and 'CONTINUE WITH ORCID' (ORCID icon). The 'REGISTER' option is also present in the modal header.



Downloading - options

Once you are logged in, **click the download button** at the top right above the data table. You will get an option to download three different things. We will be using the **"Simple" CSV** option which provides a large comma-separated spreadsheet with occurrence records as rows and different metadata for each record as columns...

The "Darwin Core Archive" format you might remember from our earlier discussions on biodiversity data formats, but we will not be using this now. The "Species List" is exactly what it says and can be useful for checklists or for checking the taxonomy of your occurrences against different taxonomic databases, but this can also be achieved using the data we will be using.

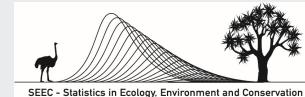
The screenshot shows the GBIF Occurrence search interface. The URL in the address bar is gbif.org/occurrence/download?country=ZA&taxon_key=3073&occurrence_status=present. The main content area displays a table of occurrence data with 190,131 results. On the right side, there are three download options:

Format	Raw data	Interpreted data	Multimedia	Coordinates	Format	Estimated data size
SIMPLE	X	✓	X	✓ (if available)	Tab-delimited CSV ⓘ	102 MB (23 MB zipped for download)
DARWIN CORE ARCHIVE	✓	✓	✓ (links)	✓ (if available)	Tab-delimited CSV ⓘ	313 MB (69 MB zipped for download)
SPECIES LIST	X	✓	X	X	Tab-delimited CSV ⓘ	

Below the download options, a "DOWNLOAD REPORT" section provides summary statistics:

- Total: 190,131
- Licence: CC BY-NC 4.0
- Year range: 1773–2021
- With year: 85 %
- With coordinates: 55 %
- With taxon match: 100 %

At the bottom left, a sidebar shows filters for occurrence status (Present), scientific name (Poaceae), and location (South Africa). The total count is 190,131.



Download in process

Once you have selected to download the "Simple" CSV, your download will process and you will get a screen like the one on the next page. This is important for two reasons:

- You will get a unique DOI for your download. This needs to be used in any publications in which you use the data from GBIF you have just downloaded.
- If you refresh the page, it will give you an update on the status of your download. Usually though you just wait until you get an email informing you that your download is ready.

Download x +

gbif.org/occurrence/download/0104208-210914110416597

Apps Google Bookmarks Gmail Mail - Vernon Visse... Bookmarks Google Calendar Google Scholar Project lists and intr... Google Docs Google Sheets Other bookmarks Reading list vervis

Get data How-to Tools Community About

DOWNLOAD | 11 JANUARY 2022

Under processing

DOI 10.15468/dl.7jzf7u

Running CANCEL

PLEASE USE THIS CITATION IN PUBLICATIONS

GBIF.org (11 January 2022) GBIF Occurrence Download <https://doi.org/10.15468/dl.7jzf7u>

Copy

FILTER APPLIED 11 JANUARY 2022 RERUN QUERY

The download has been started and is currently being processed.

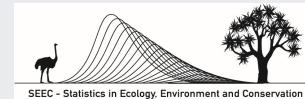
Please expect up to 3 hours for the download to complete. Most downloads will complete within 15 minutes.

A notification email with a link to download the results will be sent to the following address once ready: vervis@gmail.com

Licence: Unspecified

Make sure to read the [data user agreement](#) and [citation guidelines](#).

API



Download

Once you receive an email saying your download is ready, you can download the data to your hard drive. It is good practice to have a set of folders for any project you are working on for raw data, like this, which I usually just call "Data" and another folder for edited data and results, which I usually call "Outputs".

Using R to do GBIF downloads

Now that you have an idea of how to manually download data from GBIF, we are going to look at how to do this in R using the package `rgbif`. First we will install the package and then learn to how to search for a taxon to get a GBIF key, which is needed to do a download.

```
# Install the rgbif package. You need only do this once, i.e. the first time
# you run this code install.packages('rgbif')

# Load the packages we will need
library(rgbif)
library(dplyr) #A useful package for data wrangling

# Use the name_suggest() function to find the best species name match for
# 'Pentameris barbata'
grassName = name_suggest(q = "Pentameris barbata", rank = "species")

# What is in our grassName object?
names(grassName)

## [1] "data"      "hierarchy"

# Take a look at the data
grassName$data

## # A tibble: 1 × 3
##       key canonicalName    rank
##   <int> <chr>           <chr>
## 1 4146807 Pentameris barbata SPECIES
```



rgbif approaches to downloading

GBIF provides two ways to get occurrence data:

1. By searching for occurrences using `occ_search()`,
 2. By downloading occurrences using the numerous `occ_download*` functions.
- `occ_search()` is better for smaller data requests, while `occ_download*` functions are better for larger data requests.

Small downloads

Let's take a look at a small dataset query using `occ_search()`

```
# Let's do our search using the taxonKey we found using the name_suggest()
# function
P_barbata = occ_search(taxonKey = grassName$data$key)

# We can look at the metadata of this object using
P_barbata$meta

# We could achieve the same thing, but without first checking the taxon name,
# using:
P_barbata = occ_search(scientificName = "Pentameris barbata")

# To access the actual data you can use:
P_barbata$data
```

Some things to note about using `occ_search()` are (1) that you do not get a DOI for your download and (2) you are limited to a maximum of 100,000 records. You need a DOI if you want to publish using these data, but it is possible to get this using the `derived_dataset()` function. However, we will rather use the `occ_download*`() approach from here on in.

Large downloads

Let's see how to do a large download using our earlier example of downloading all *Pentameris* records in South Africa.

```
# First get the taxon key
taxonKey = name_suggest(q = 'Pentameris', rank = "genus")

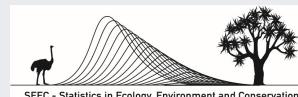
# Set up your username etc., which will be needed later
# user = "" # GBIF user name
# pwd = "" # GBIF password
# email = "" # Your email

# Set up the download. This query is sent to GBIF, just as you did manually and then we need to import the
download after this.
PentamerisInSADownload = occ_download(
  pred_in("taxonKey", taxonKey$data$key), # Set the taxonKey
  pred("hasCoordinate", TRUE), # Select only records with coordinates
  pred("country", "ZA"), # Select only records in South Africa
  format = "SIMPLE_CSV", # Get simple CSV format
  user = user, pwd = pwd, email = email # Give rgbif your credentials for GBIF
)

# We can now look at the download key
PentamerisInSADownload

# Let us take a look and see if our download is ready and see metadata associated with it
occ_download_meta(key = "0081229-230224095556074")

# Download the dataset from GBIF to a folder you choose and import it into R
PentamerisInSA = occ_download_get(key = "0081229-230224095556074", path = "../Data/", overwrite = TRUE) %>%
#Download the data
  occ_download_import(PentamerisInSA, na.strings = c("", NA), quote = "") #Import data into R
```



Importing very large downloads

Using `occ_download_import()` works for large but not VERY large datasets. If your download is larger than say a million records, you might be better off just running the `occ_download_get()` part of the code and then reading the data in using a "large data" package, such as `data.table`.

Before you do this, you can unzip the GBIF download using e.g., 7zip, or directly from within R and then read the CSV file into R, like below.

```
unzip("../Data/0081229-230224095556074.zip", exdir="../Data")  
  
#Read CSV data into R  
# install.packages(data.table) #First install data.table, if you do not already have it  
library(data.table)  
dat = fread("../Data/0081229-230224095556074.csv", data.table = FALSE, fill = F, encoding = "UTF-8", quote="")
```

Take a look at our data

```
names(dat)
```

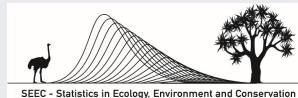
```
## [1] "gbifID"                                "datasetKey"  
## [3] "occurrenceID"                            "kingdom"  
## [5] "phylum"                                 "class"  
## [7] "order"                                  "family"  
## [9] "genus"                                   "species"  
## [11] "infraspecificEpithet"                   "taxonRank"  
## [13] "scientificName"                          "verbatimScientificName"  
## [15] "verbatimScientificNameAuthorship"        "countryCode"  
## [17] "locality"                               "stateProvince"  
## [19] "occurrenceStatus"                        "individualCount"  
## [21] "publishingOrgKey"                        "decimalLatitude"  
## [23] "decimalLongitude"                        "coordinateUncertaintyInMeters"  
## [25] "coordinatePrecision"                     "elevation"  
## [27] "elevationAccuracy"                      "depth"  
## [29] "depthAccuracy"                           "eventDate"  
## [31] "day"                                     "month"  
## [33] "year"                                    "taxonKey"  
## [35] "speciesKey"                             "basisOfRecord"  
## [37] "institutionCode"                         "collectionCode"  
## [39] "catalogNumber"                           "recordNumber"  
## [41] "identifiedBy"                            "dateIdentified"  
## [43] "license"                                 "rightsHolder"  
## [45] "recordedBy"                             "typeStatus"  
## [47] "establishmentMeans"                     "lastInterpreted"  
## [49] "mediaType"                              "issue"
```



Take a look at our data

```
head(dat)
```

```
##      gbifID          datasetKey
## 1 912558605 cd6e21c8-9e8a-493a-8a76-fbf7862069e5
## 2 912558604 cd6e21c8-9e8a-493a-8a76-fbf7862069e5
## 3 912558594 cd6e21c8-9e8a-493a-8a76-fbf7862069e5
## 4 912558592 cd6e21c8-9e8a-493a-8a76-fbf7862069e5
## 5 912558591 cd6e21c8-9e8a-493a-8a76-fbf7862069e5
## 6 912558587 cd6e21c8-9e8a-493a-8a76-fbf7862069e5
##           occurrenceID kingdom      phylum      class
## 1 http://specimens.kew.org/herbarium/K000719928 Plantae Tracheophyta Liliopsida
## 2 http://specimens.kew.org/herbarium/K000719938 Plantae Tracheophyta Liliopsida
## 3 http://specimens.kew.org/herbarium/K000719932 Plantae Tracheophyta Liliopsida
## 4 http://specimens.kew.org/herbarium/K000719940 Plantae Tracheophyta Liliopsida
## 5 http://specimens.kew.org/herbarium/K000719919 Plantae Tracheophyta Liliopsida
## 6 http://specimens.kew.org/herbarium/K000719925 Plantae Tracheophyta Liliopsida
##   order family   genus      species infraspecificEpithet taxonRank
## 1 Poales Poaceae Pentameris   Pentameris dregeana                   SPECIES
## 2 Poales Poaceae Pentameris   Pentameris pusilla                   SPECIES
## 3 Poales Poaceae Pentameris   Pentameris glacialis                  SPECIES
## 4 Poales Poaceae Pentameris   Pentameris pusilla                   SPECIES
## 5 Poales Poaceae Pentameris   Pentameris dregeana                  SPECIES
## 6 Poales Poaceae Pentameris   Pentameris dregeana                  SPECIES
##           scientificName      verbatimScientificName
## 1      Pentameris dregeana Staph      Pentameris dregeana Staph
## 2 Poagrostis pusilla (Nees) Staph Poagrostis pusilla (Nees) Staph
## 3 Pentameris glacialis N.P.Barker Pentameris glacialis N.P.Barker
## 4 Poagrostis pusilla (Nees) Staph Poagrostis pusilla (Nees) Staph
## 5      Pentameris dregeana Staph      Pentameris dregeana Staph
## 6      Pentameris dregeana Staph      Pentameris dregeana Staph
##   verbatimScientificNameAuthorship countryCode
## 1                               Staph        ZA
## 2 (Nees) Staph        ZA
## 3      N.P.Barker        ZA
## 4 (Nees) Staph        ZA
## 5                               Staph        ZA
```



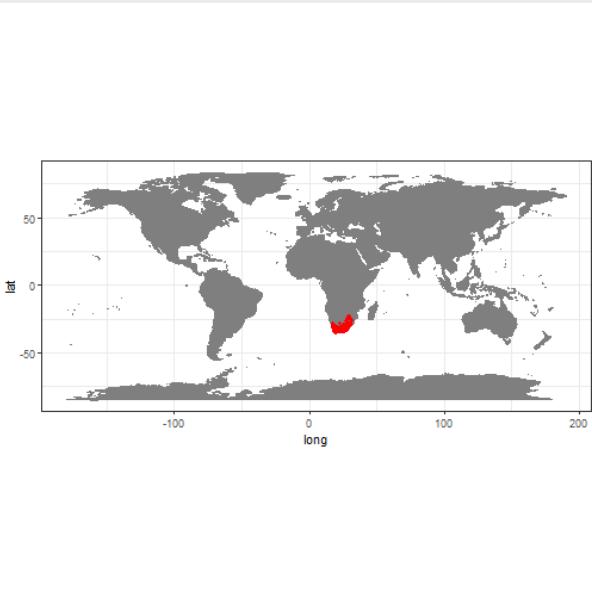
SANBI
Biodiversity for Life

South African National Biodiversity Institute

Make a map of our occurrences

```
# install.packages(ggplot2) #First install ggplot2, if you do not already have it
library(ggplot2) #Needed for making plots
#Get a world map that is just grey in colour
wm = borders("world", colour="gray50", fill="gray50")

#Plot dataset
ggplot(dat=dat) +
  coord_fixed() + #Specify coordinate system
  wm + #Add world map
  geom_point(aes(x = decimalLongitude, y = decimalLatitude), colour = 'red', size = 1) + #Add points for
occurrences
  theme_bw() #Specify a ggplot theme
```

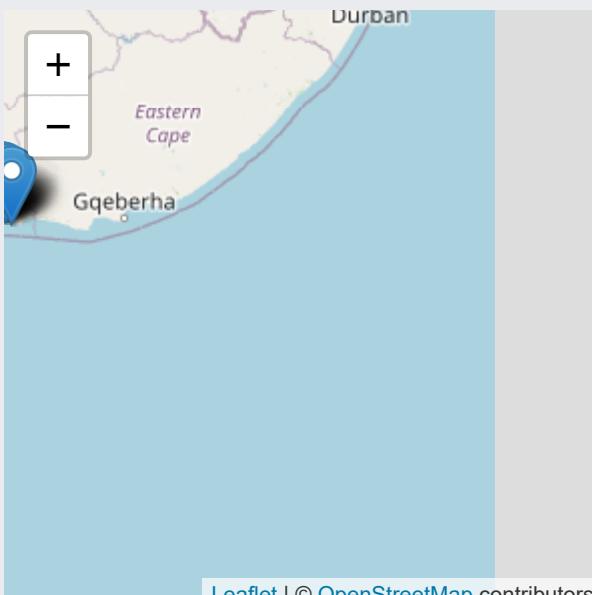


Interactive map option

```
# install.packages(leaflet) #First install leaflet, if you do not already have it
library(leaflet) #Needed for making plots

P_barbata = dat %>%
  filter(species=="Pentameris barbata") %>% #Select only Pentameris barbata occurrences
  filter(!is.na(decimalLongitude)) #Remove occurrences with no coordinates

leafP_barbata = leaflet::leaflet() %>%
  leaflet::addProviderTiles(providers$OpenStreetMap) %>%
  leaflet::addMarkers(
    lng = P_barbata$decimalLongitude,
    lat = P_barbata$decimalLatitude,
    data = P_barbata,
    popup = ~as.character(P_barbata$gbifID)
  )
leafP_barbata
```



Leaflet | © OpenStreetMap contributors



SEEC - Statistics in Ecology, Environment and Conservation



SANBI 
Biodiversity for Life
South African National Biodiversity Institute

Now you have data, but need to interrogate and clean it next...



SEEC - Statistics in Ecology, Environment and Conservation



South African National Biodiversity Institute