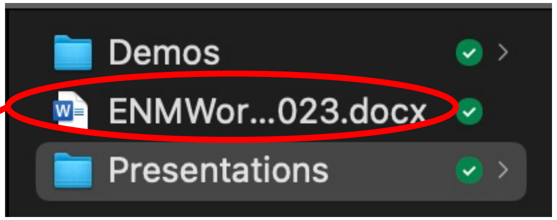
# R-based Data Downloads

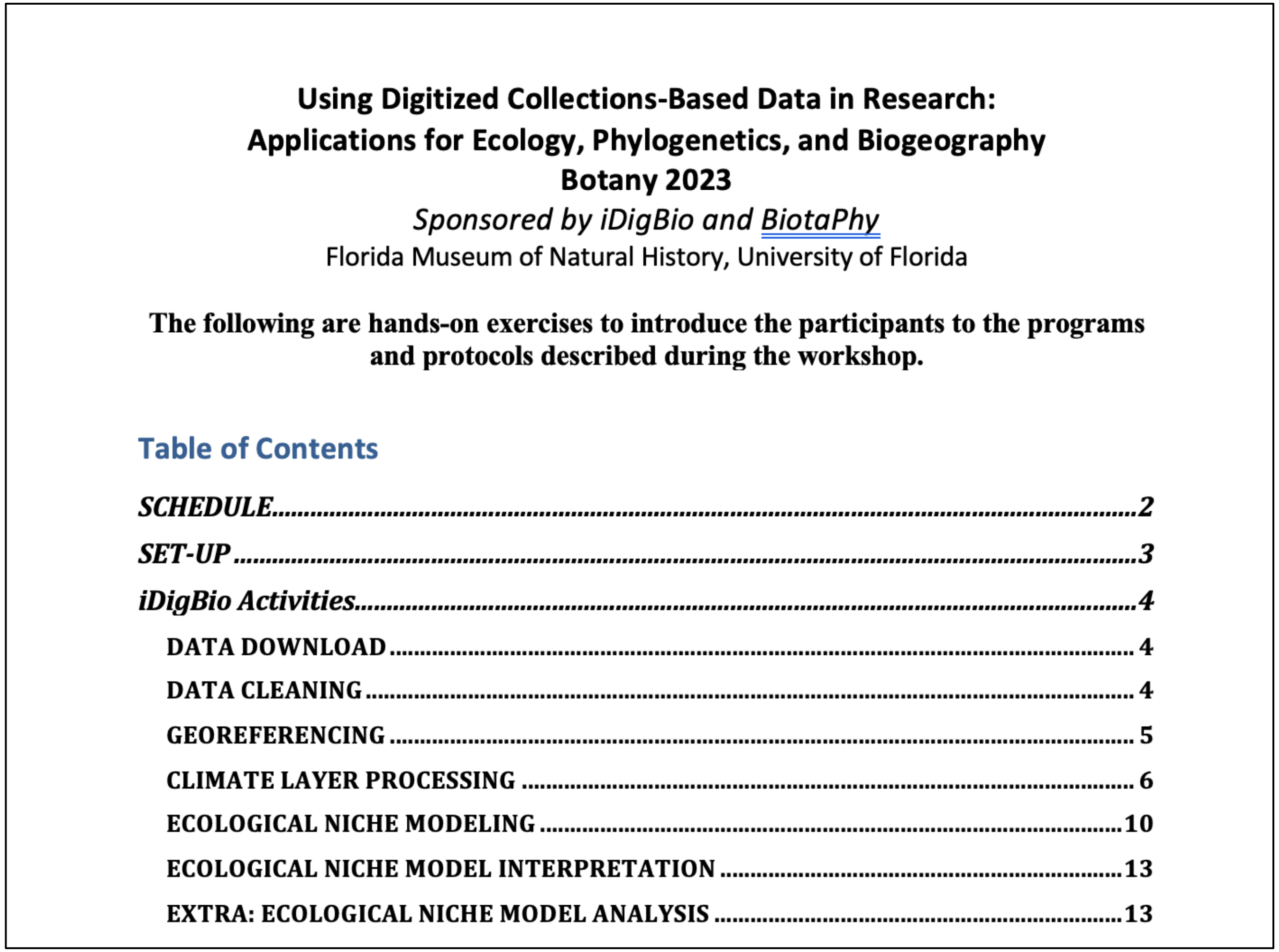
Shelly Gaynor
University of Florida

iDigBio
Integrated Digitized Biocollections

# Workshop setup



Workshop Guide

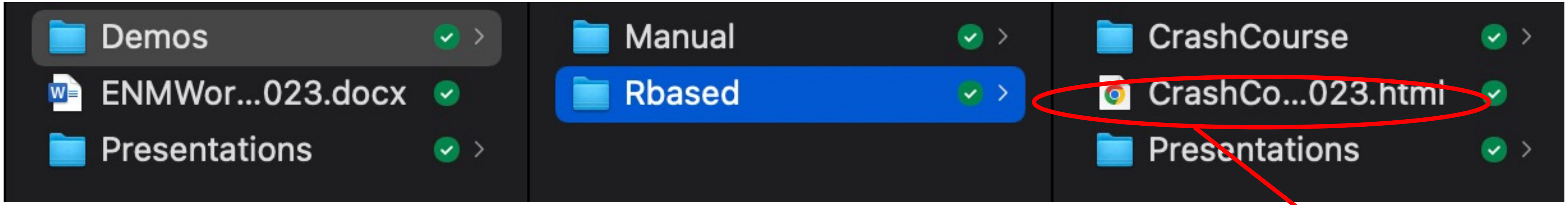


**Using Digitized Collections-Based Data in Research:**
**Applications for Ecology, Phylogenetics, and Biogeography**
**Botany 2023**
*Sponsored by iDigBio and BiotaPhy*
Florida Museum of Natural History, University of Florida

**The following are hands-on exercises to introduce the participants to the programs and protocols described during the workshop.**

## Table of Contents

# Workshop setup



Demos

ENMWor...023.docx

Presentations

Manual

Rbased

CrashCourse

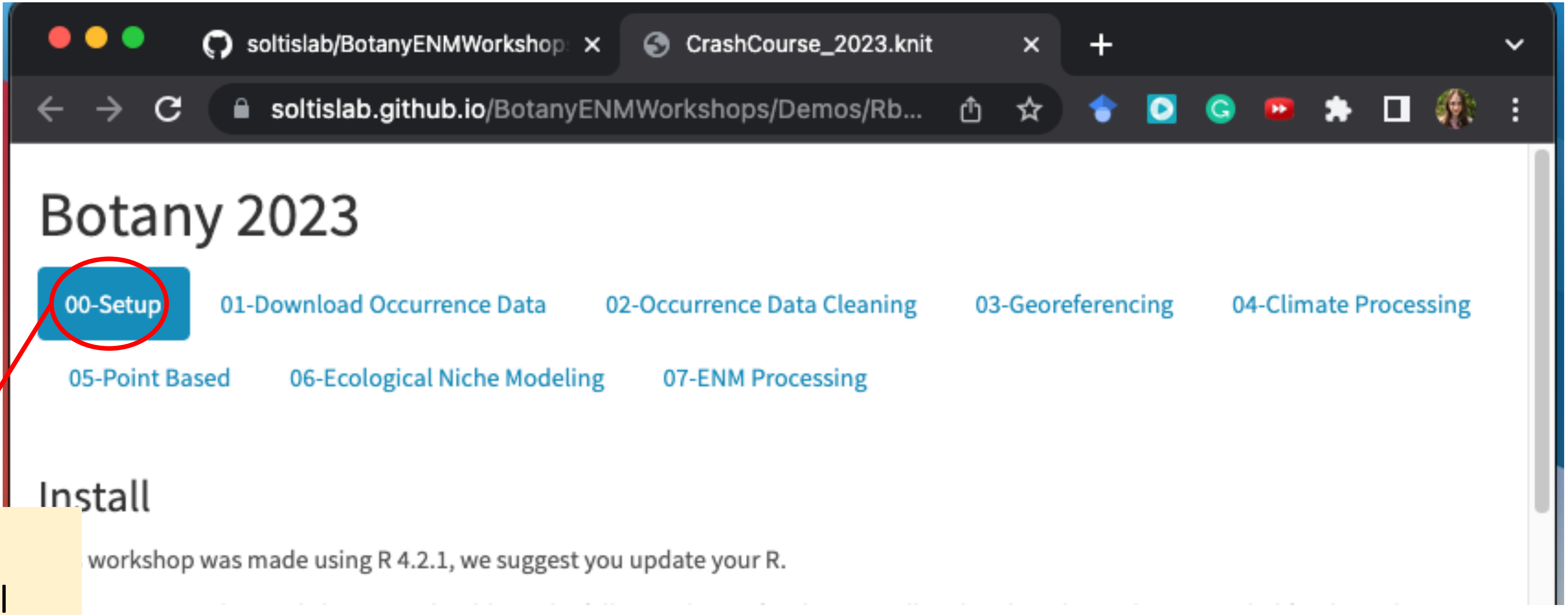
CrashCo...023.html

Presentations

HTML Version of all R scripts

# Workshop setup

https://tinyurl.com/4nk7zcws


soltislab/BotanyENMWorkshop
CrashCourse_2023.knit
soltislab.github.io/BotanyENMWorkshops/Demos/Rb...

Botany 2023

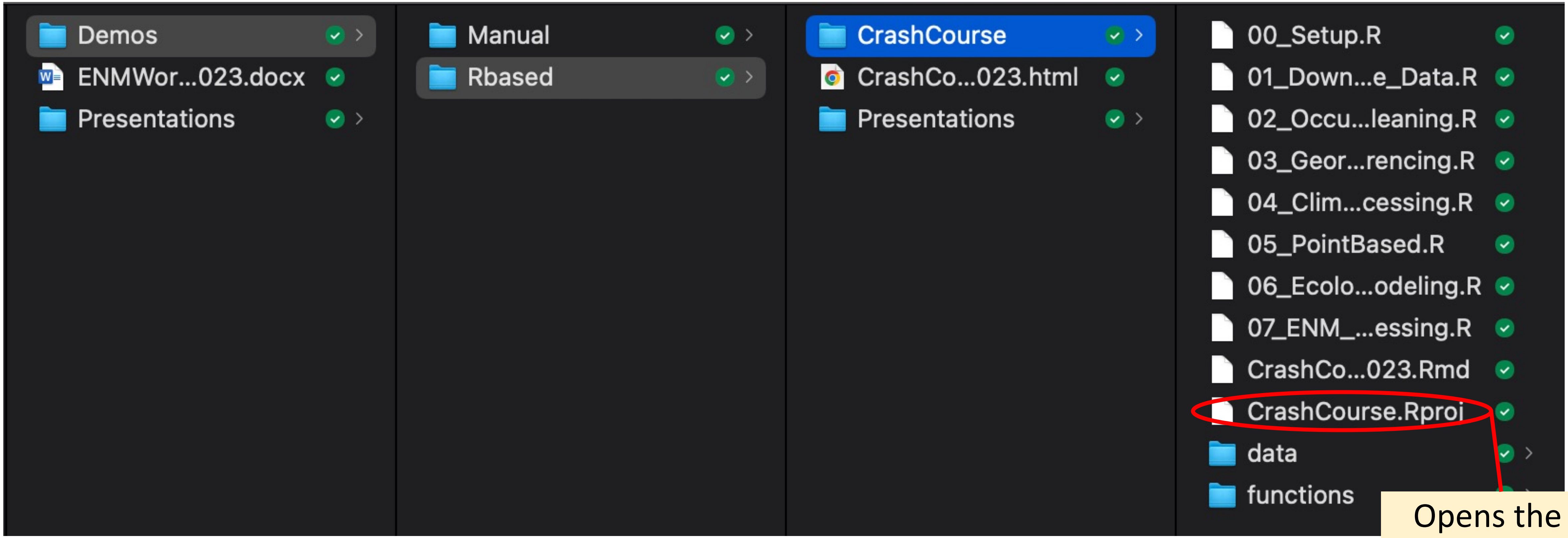
00-Setup
01-Download Occurrence Data
02-Occurrence Data Cleaning
03-Georeferencing
04-Climate Processing
05-Point Based
06-Ecological Niche Modeling
07-ENM Processing

Install

workshop was made using R 4.2.1, we suggest you update your R.

Contains additional troubleshooting steps

# Workshop setup



Demos

ENMWor...023.docx

Presentations

Manual

Rbased

CrashCourse

CrashCo...023.html

Presentations

00_Setup.R

01_Down...e_Data.R

02_Occu...leaning.R

03_Geor...rencing.R

04_Clim...cessing.R

05_PointBased.R

06_Ecolo...odeling.R

07_ENM_...essing.R

CrashCo...023.Rmd

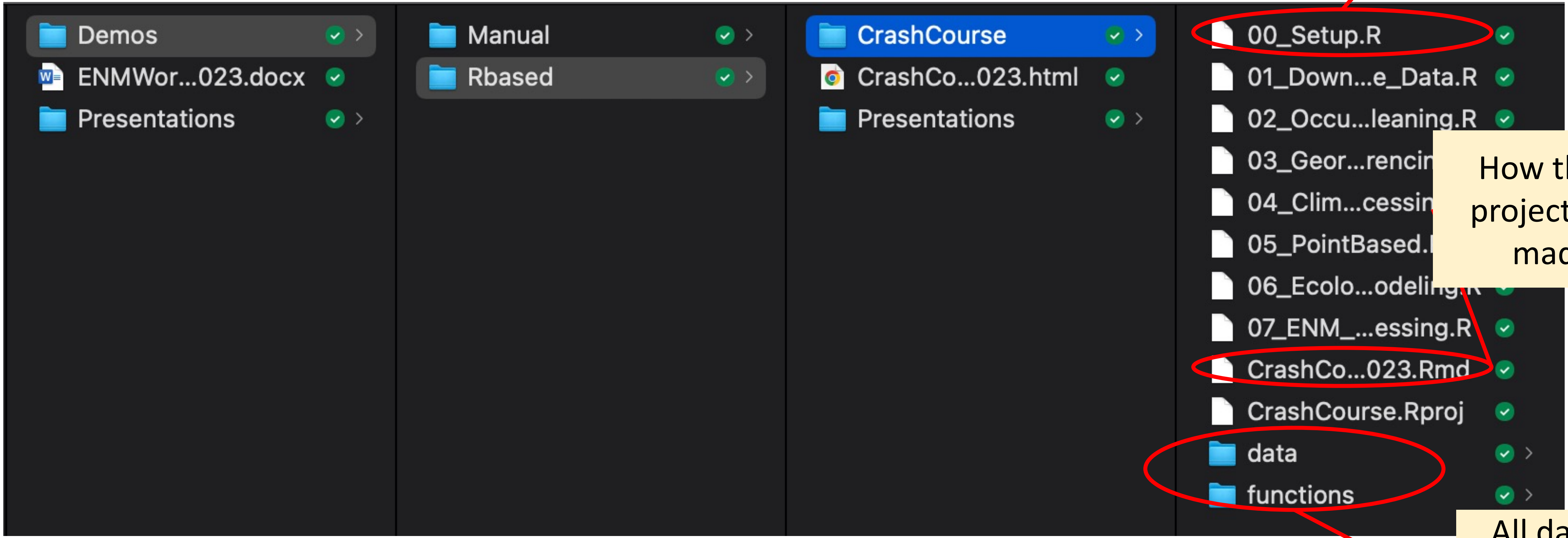
CrashCourse.Rproj

data

functions

Opens the R project

# Workshop setup

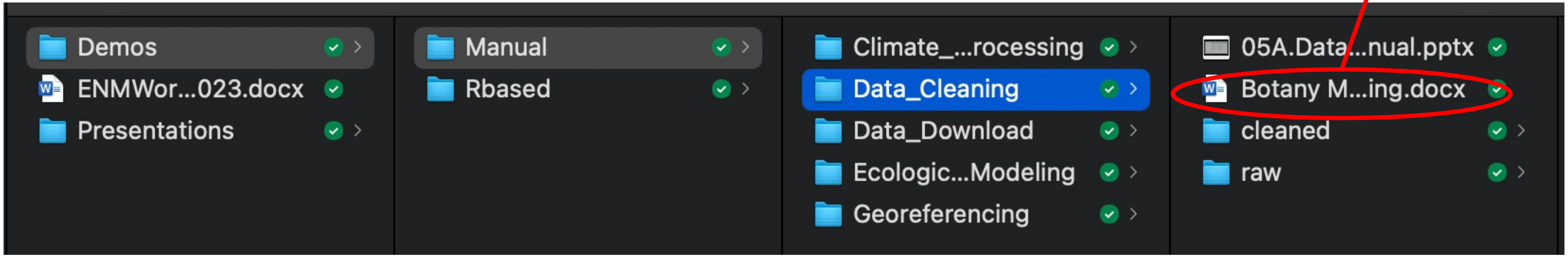


Run before the workshop started!

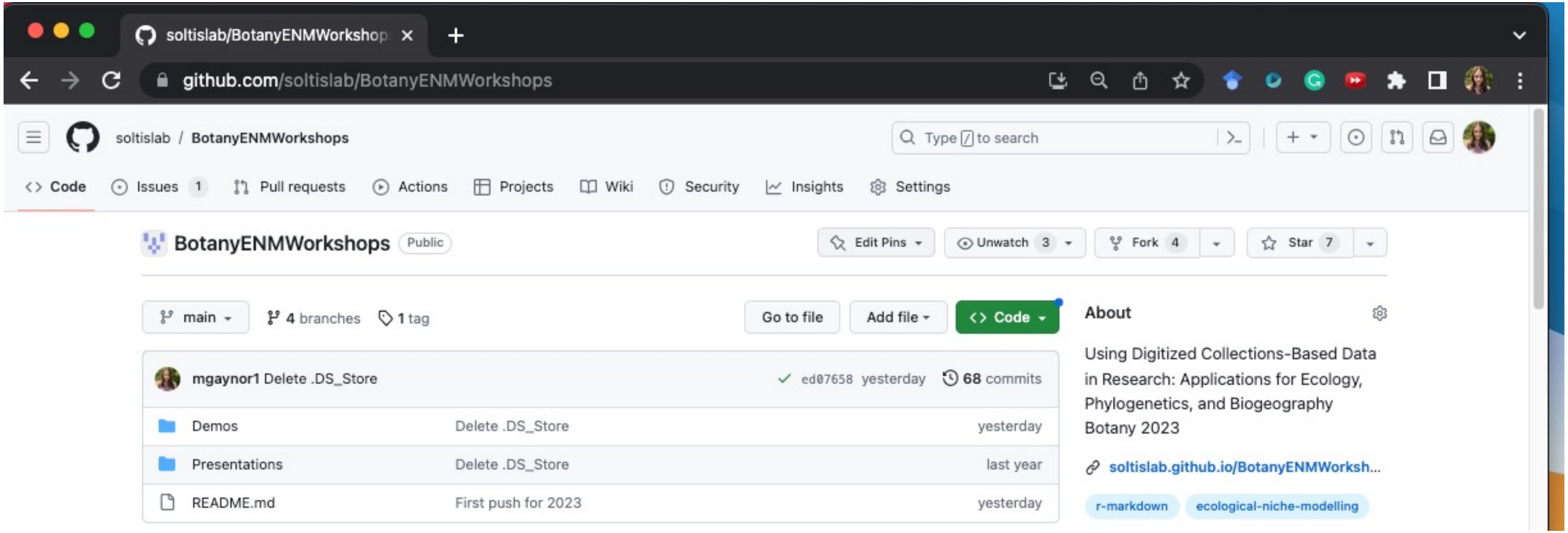
How the R project was made

All data and functions needed for these scripts

# Workshop setup



Instructions



mgaynor1 Delete .DS_Store ✓ ed07658 yesterday 68 commits

| | | |
|---|---|---|
| Demos | Delete .DS_Store | yesterday |
| Presentations | Delete .DS_Store | last year |
| README.md | First push for 2023 | yesterday |

**About**

Using Digitized Collections-Based Data in Research: Applications for Ecology, Phylogenetics, and Biogeography Botany 2023

soltislab.github.io/BotanyENMWorksh...

r-markdown ecological-niche-modelling

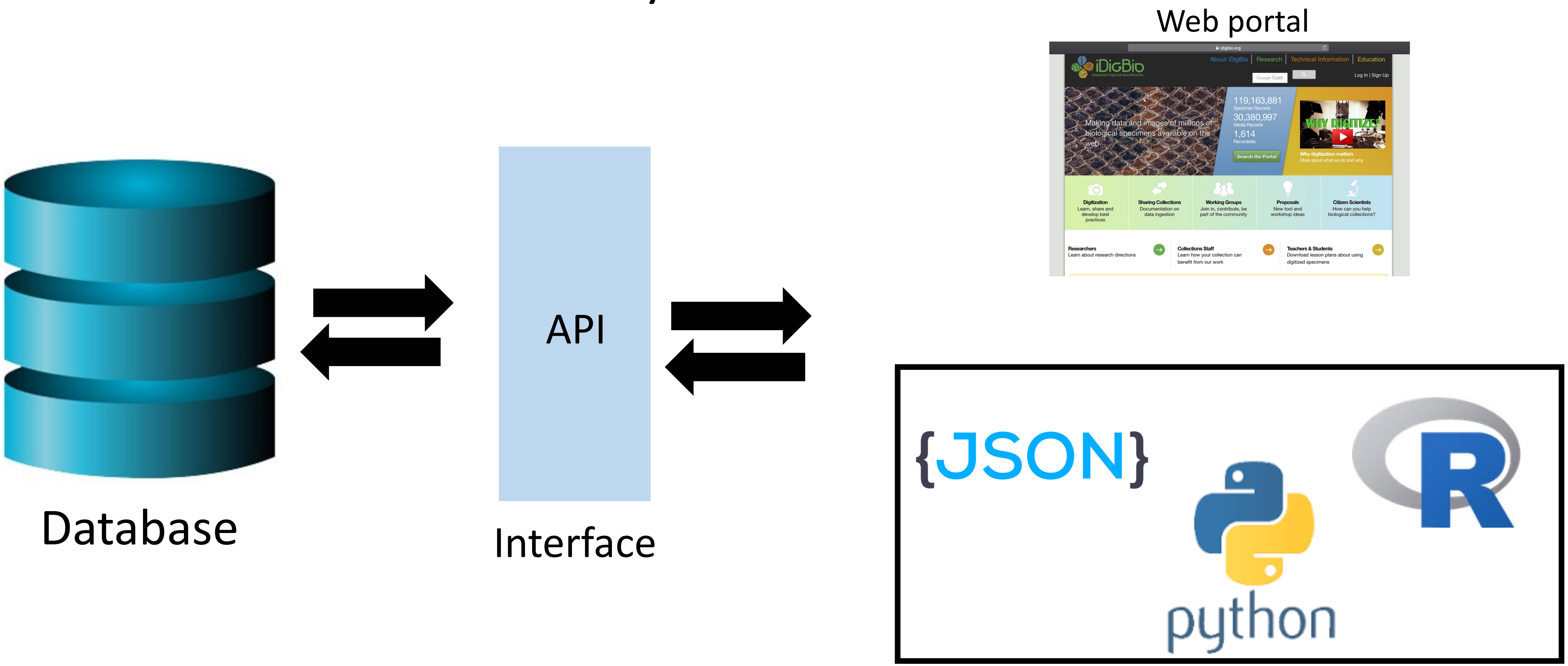
# 01-Data Download

University of Florida

iDigBio
Integrated Digitized Biocollections

# API = Application Programming Interface

- Allows users to interact with a system



Web portal

Database

API

Interface

Making data and images of millions of biological specimens available on the web

119,163,881 Specimen Records

30,380,997 Media Records

1,614 Recordsets

Why digitization matters
More about what we do and why

**Digitization** Learn, share and develop best practices

**Sharing Collections** Documentation on data ingestion

**Working Groups** Join in, contribute, be part of the community

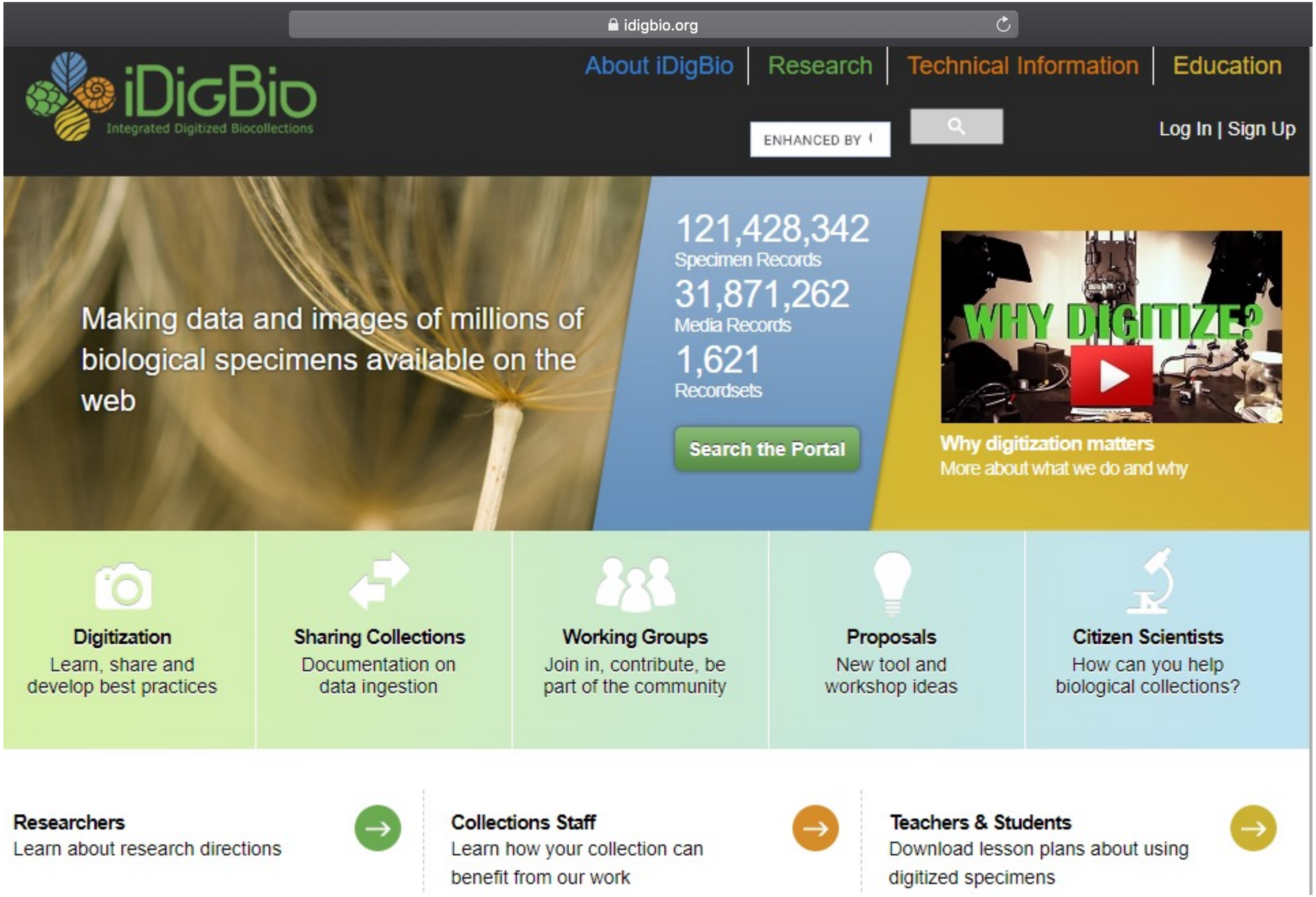
**Proposals** New tool and workshop ideas

**Citizen Scientists** How can you help biological collections?

**Researchers** Learn about research directions

**Collections Staff** Learn how your collection can benefit from our work

**Teachers & Students** Download lesson plans about using digitized specimens



Making data and images of millions of biological specimens available on the web

121,428,342 Specimen Records

31,871,262 Media Records

1,621 Recordsets

**Why digitization matters**
More about what we do and why

**Digitization**
Learn, share and develop best practices

**Sharing Collections**
Documentation on data ingestion

**Working Groups**
Join in, contribute, be part of the community

**Proposals**
New tool and workshop ideas

**Citizen Scientists**
How can you help biological collections?

**Researchers**
Learn about research directions

**Collections Staff**
Learn how your collection can benefit from our work

**Teachers & Students**
Download lesson plans about using digitized specimens

# iDigBio API

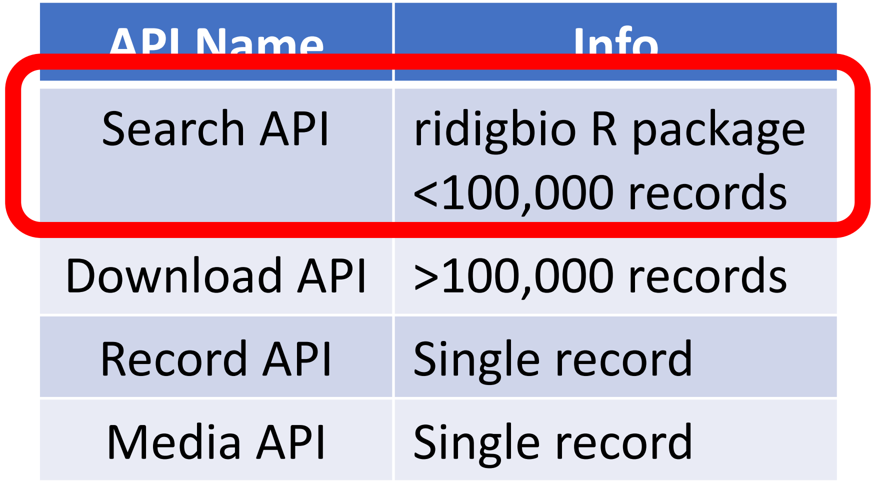
- Multiple ways to access the API:

| API Name | Info |
|---|---|
| Search API | ridigbio R package<br><100,000 records |
| Download API | >100,000 records |
| Record API | Single record |
| Media API | Single record |

biodiversity-specimen-data/**specimen-data-use-case**

# iDigBio API

- Multiple ways to access the API:



| API Name | Info |
|---|---|
| Search API | ridigbio R package<br><100,000 records |
| Download API | >100,000 records |
| Record API | Single record |
| Media API | Single record |

biodiversity-specimen-data/**specimen-data-use-case**

# GBIF API

- Multiple ways to access the API:

| API Name | Info |
|---|---|
| Registry API | Create, edit, update and search for information about datasets |
| Species API | Taxonomy API |
| Occurrence API | Record API |
| Maps API | Show maps of GBIF |
| News API | Search papers published using GBIF |

GBIF Global Biodiversity Information Facility

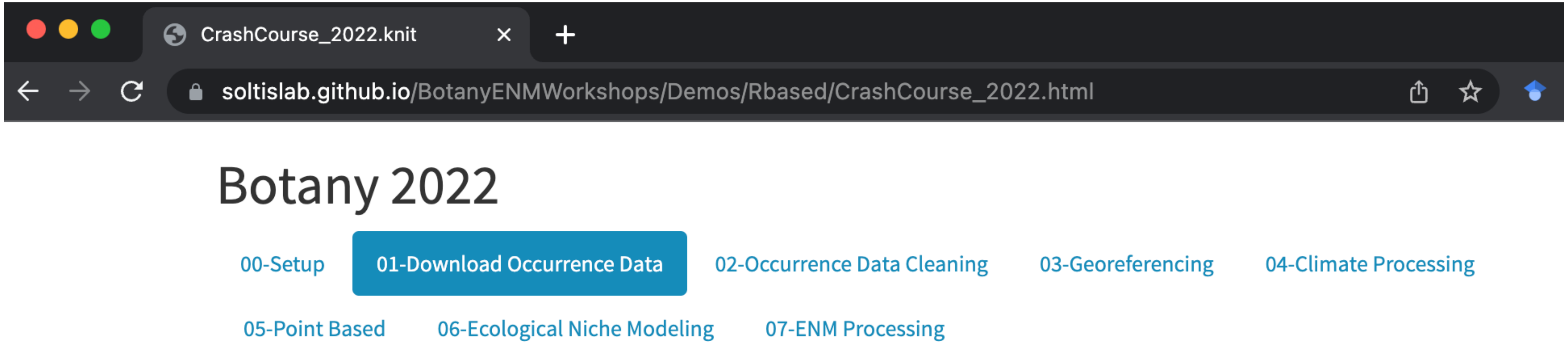
https://www.gbif.org/developer/summary

# R based

https://tinyurl.com/msdurpk

"Demo/Rbased/CrashCourse/CrashCourse.Rproj"

- Navigate to 01_Download_Occurence_Data.R

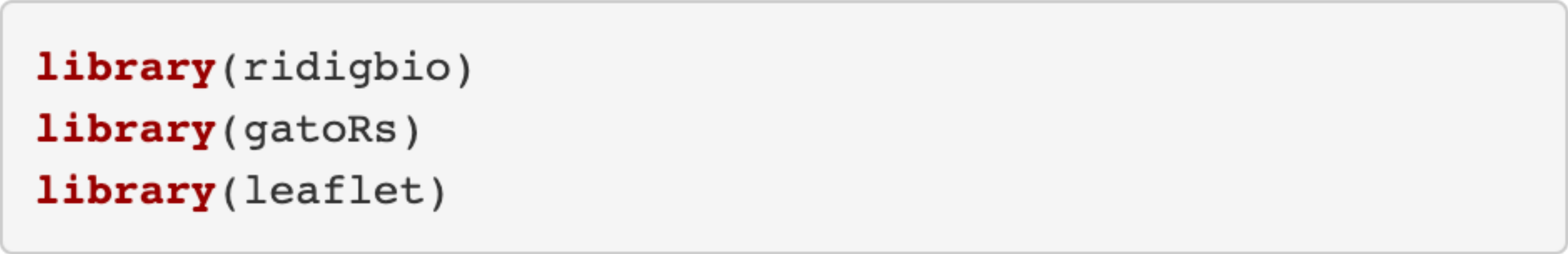


CrashCourse_2022.knit
soltislab.github.io/BotanyENMWorkshops/Demos/Rbased/CrashCourse_2022.html

Botany 2022

00-Setup
01-Download Occurrence Data
02-Occurrence Data Cleaning
03-Georeferencing
04-Climate Processing
05-Point Based
06-Ecological Niche Modeling
07-ENM Processing

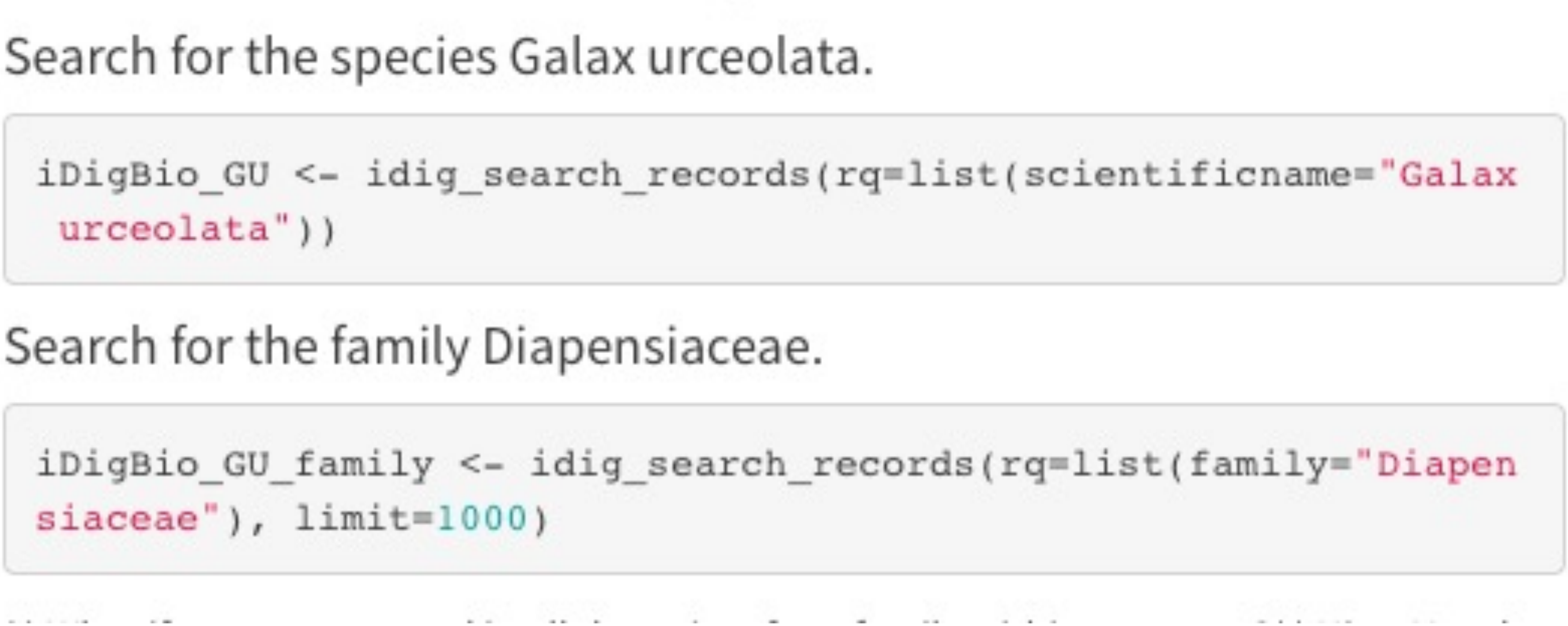
# Load Packages



```r
library(ridigbio)
library(gatoRs)
library(leaflet)
```

# Downloading data using ridigbio

- First, we are searching for the species Galax urceolata
- Next, download occurrence records for the family Diapensiaceae



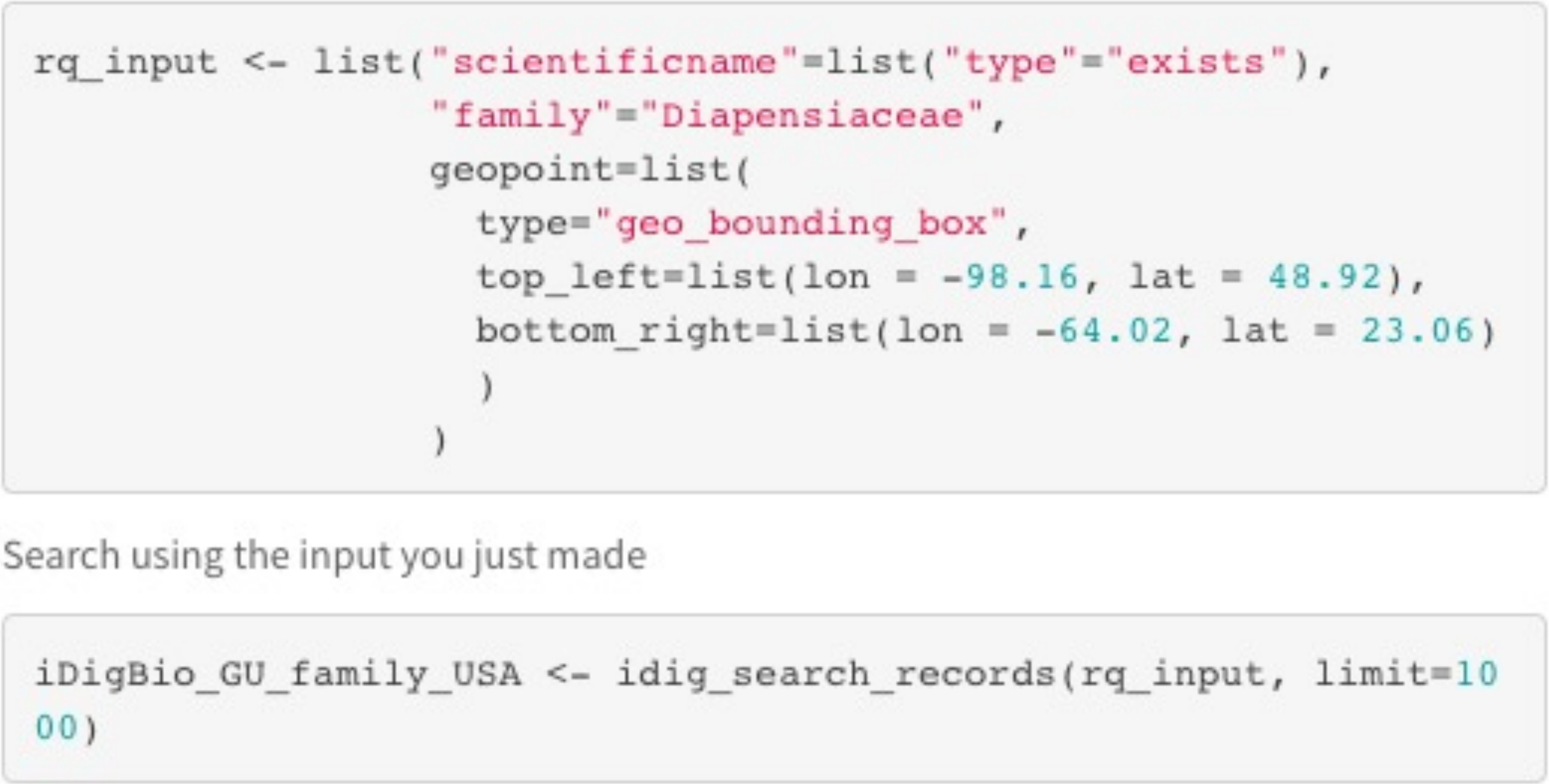
Search for the species Galax urceolata.

```
iDigBio_GU <- idig_search_records(rq=list(scientificname="Galax
 urceolata"))
```

Search for the family Diapensiaceae.

```
iDigBio_GU_family <- idig_search_records(rq=list(family="Diapen
siaceae"), limit=1000)
```

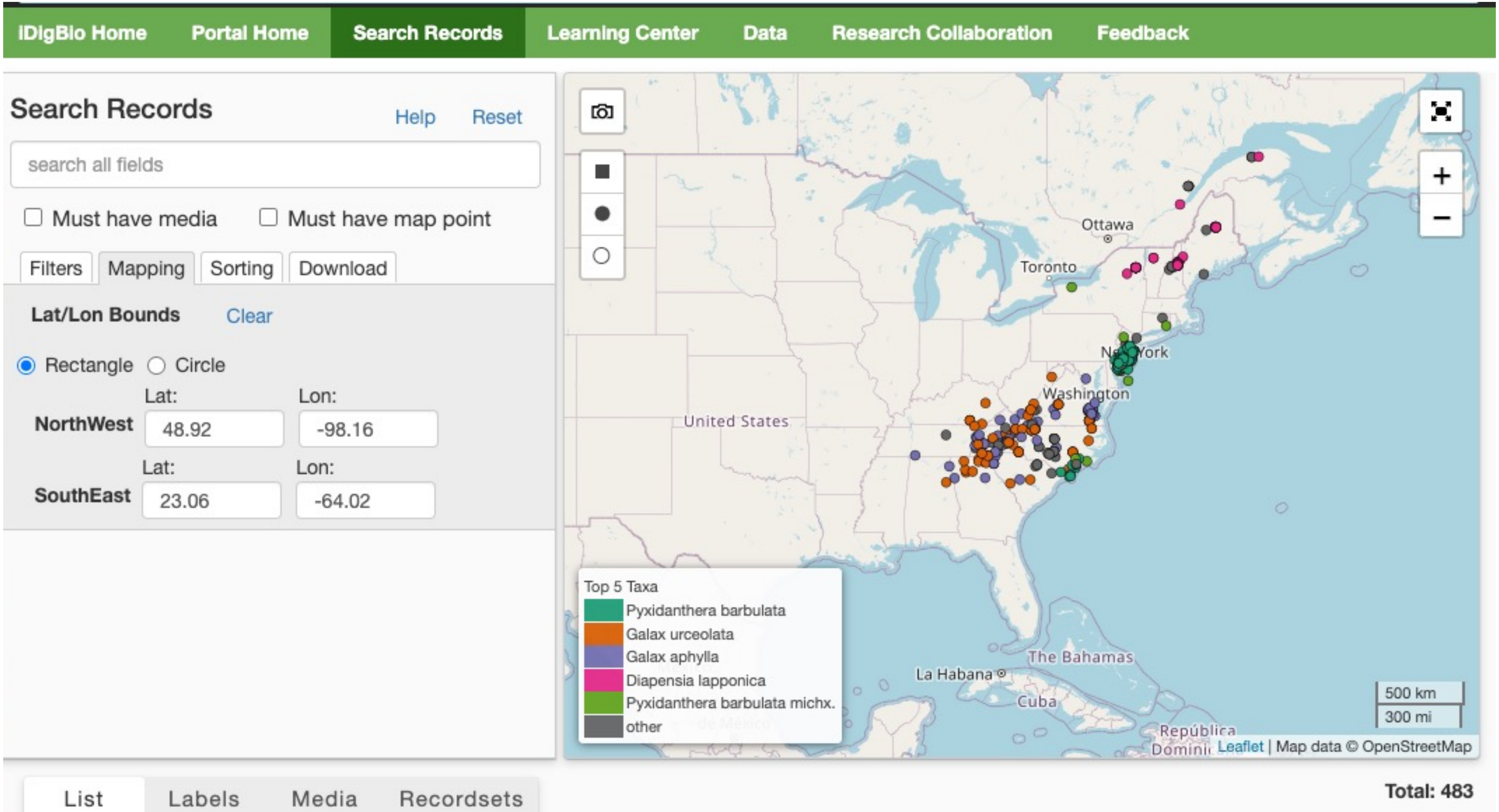
# Records only in North America



```
rq_input <- list("scientificname"=list("type"="exists"),
                 "family"="Diapensiaceae",
                 geopoint=list(
                   type="geo_bounding_box",
                   top_left=list(lon = -98.16, lat = 48.92),
                   bottom_right=list(lon = -64.02, lat = 23.06)
                   )
                 )
```

Search using the input you just made

```
iDigBio_GU_family_USA <- idig_search_records(rq_input, limit=10
00)
```



## Search Records

search all fields

☐ Must have media ☐ Must have map point

**Lat/Lon Bounds**

● Rectangle ○ Circle

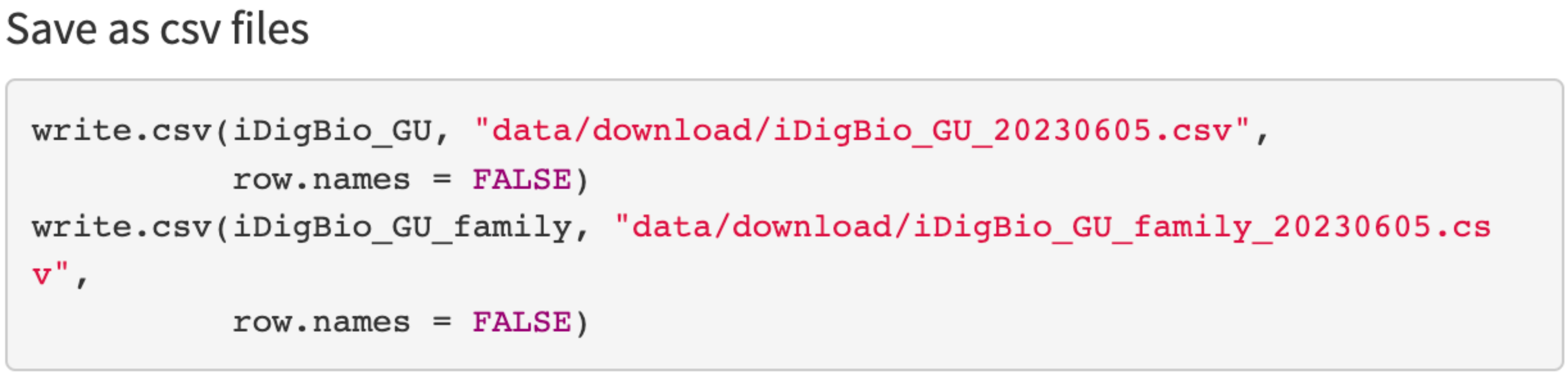
| | Lat: | Lon: |
|---|---|---|
| **NorthWest** | 48.92 | -98.16 |
| **SouthEast** | 23.06 | -64.02 |

Top 5 Taxa

- 🟩 Pyxidanthera barbulata
- 🟧 Galax urceolata
- 🟪 Galax aphylla
- 🟥 Diapensia lapponica
- 🟩 Pyxidanthera barbulata michx.
- ⬛ other

**Total: 483**

# Save as csv



Save as csv files

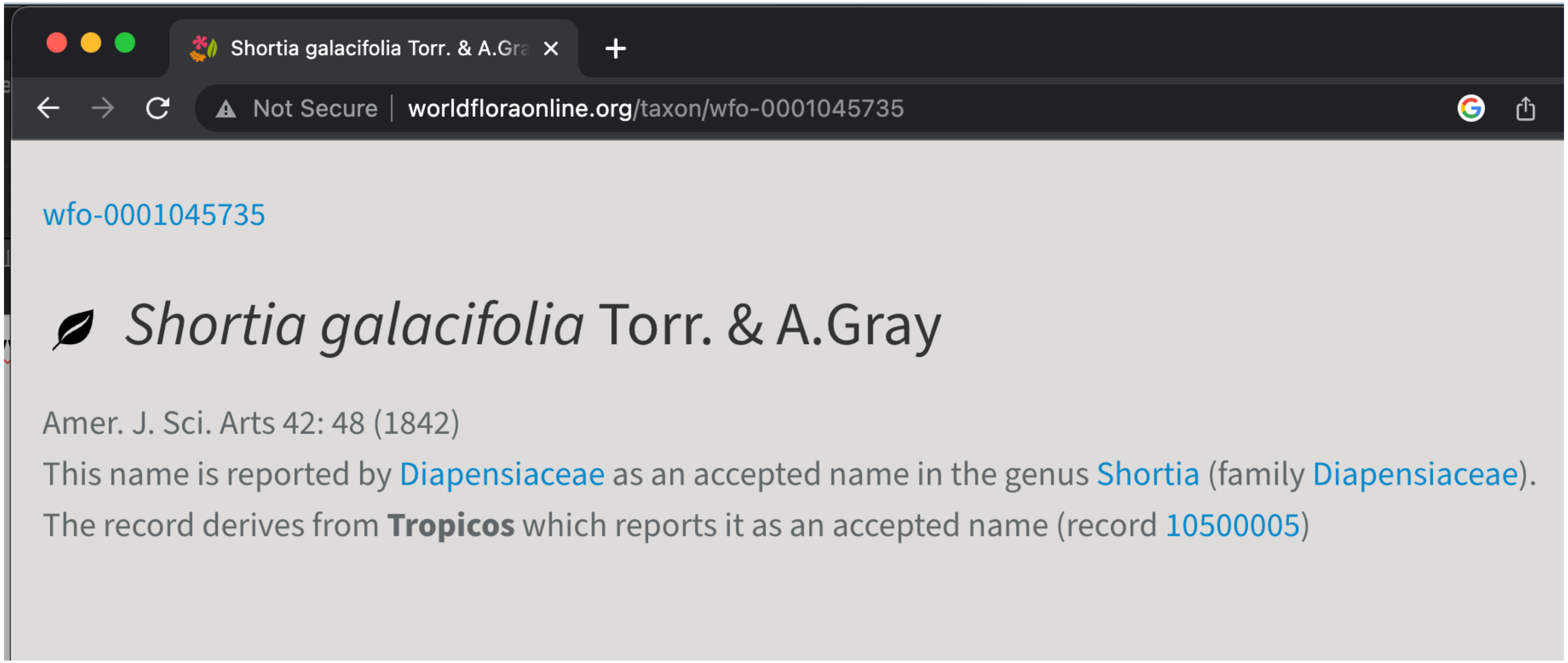
```
write.csv(iDigBio_GU, "data/download/iDigBio_GU_20230605.csv",
          row.names = FALSE)
write.csv(iDigBio_GU_family, "data/download/iDigBio_GU_family_20230605.cs
v",
          row.names = FALSE)
```

# Data download using gatoRs

Natalie Patten

- To pull data from GBIF and iDigBio for a set of synonyms
- Identifying synonyms:
  - Taxonomic Name Resolution Service
    - https://tnrs.biendata.org/
    - Used in soltislab/BotanyENMWorkshops 2020
  - R package taxize
    - 20 sources for synonyms
    - https://docs.ropensci.org/taxize/

http://www.worldfloraonline.org/



Shortia galacifolia Torr. & A.Gra ×
+
Not Secure
worldfloraonline.org/taxon/wfo-0001045735

wfo-0001045735

## *Shortia galacifolia* Torr. & A.Gray

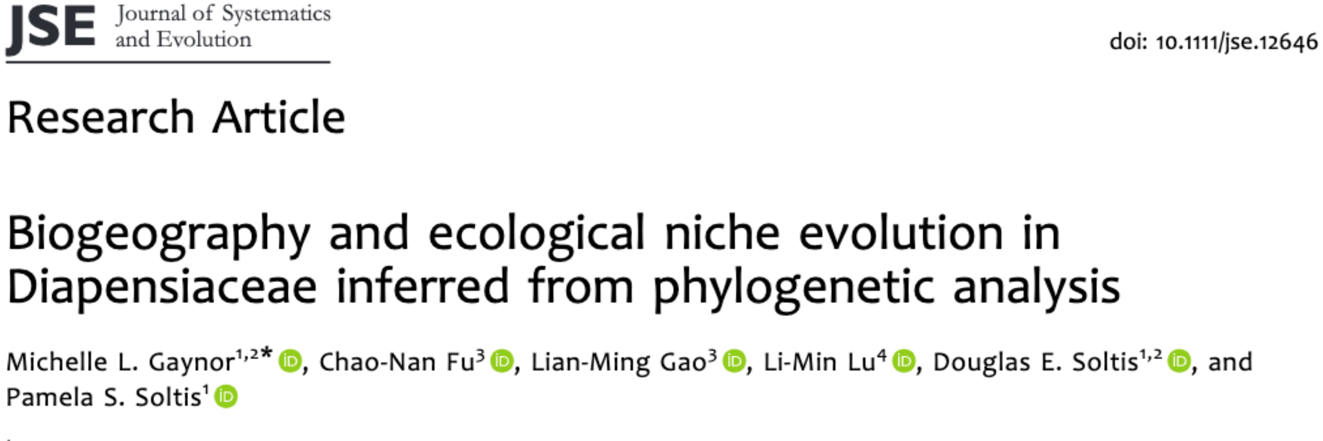
Amer. J. Sci. Arts 42: 48 (1842)

This name is reported by Diapensiaceae as an accepted name in the genus Shortia (family Diapensiaceae).

The record derives from **Tropicos** which reports it as an accepted name (record 10500005)

World Flora Online

Where these synonyms are from:



JSE Journal of Systematics and Evolution

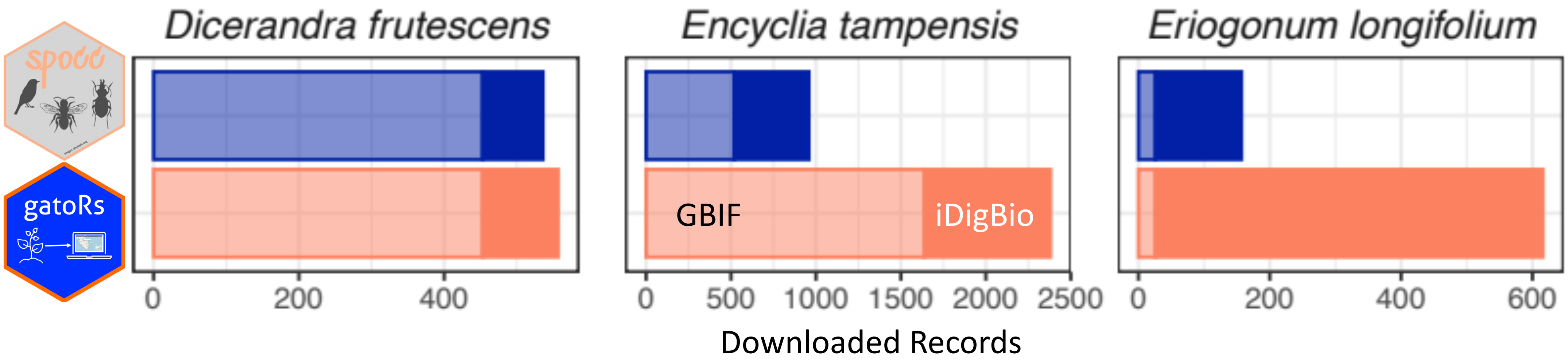
doi: 10.1111/jse.12646

Research Article

### Biogeography and ecological niche evolution in Diapensiaceae inferred from phylogenetic analysis

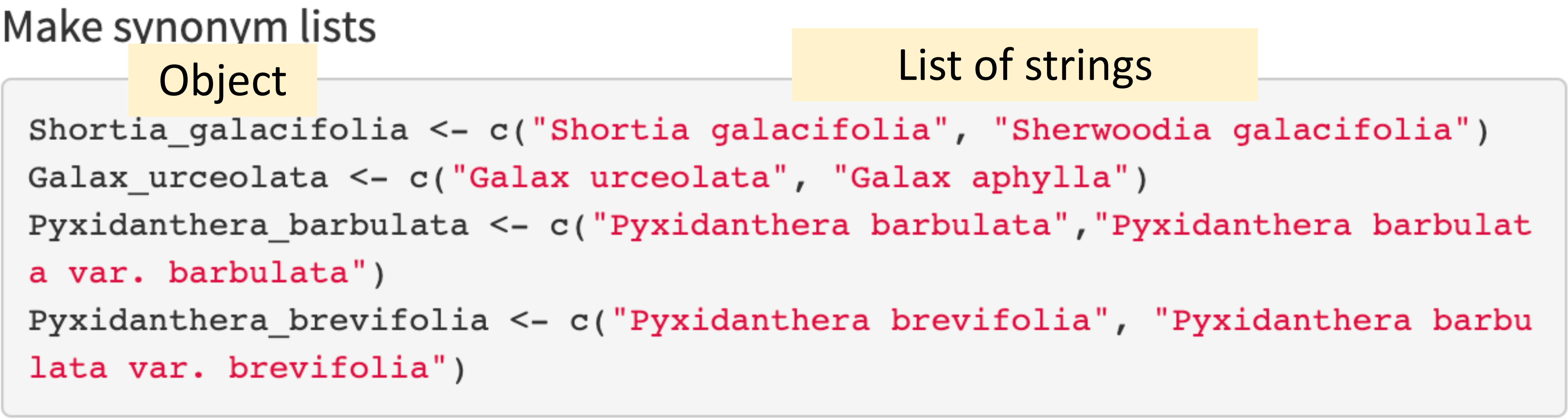
Michelle L. Gaynor[1,2]*, Chao-Nan Fu[3], Lian-Ming Gao[3], Li-Min Lu[4], Douglas E. Soltis[1,2], and Pamela S. Soltis[1]

# Why should I use gatoRs?



Patten et al. In Review. Applications in Plant Science

# Data download using gator_download



Make synonym lists

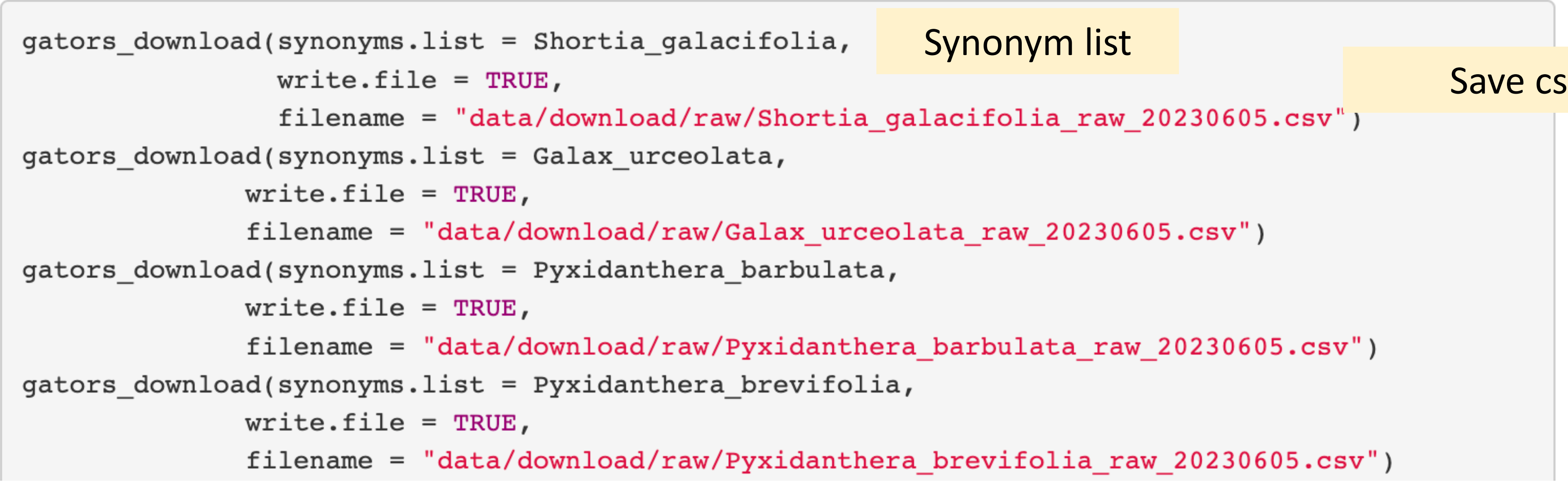
Object

List of strings

```
Shortia_galacifolia <- c("Shortia galacifolia", "Sherwoodia galacifolia")
Galax_urceolata <- c("Galax urceolata", "Galax aphylla")
Pyxidanthera_barbulata <- c("Pyxidanthera barbulata","Pyxidanthera barbulat
a var. barbulata")
Pyxidanthera_brevifolia <- c("Pyxidanthera brevifolia", "Pyxidanthera barbu
lata var. brevifolia")
```

# Data download using gators_download

,



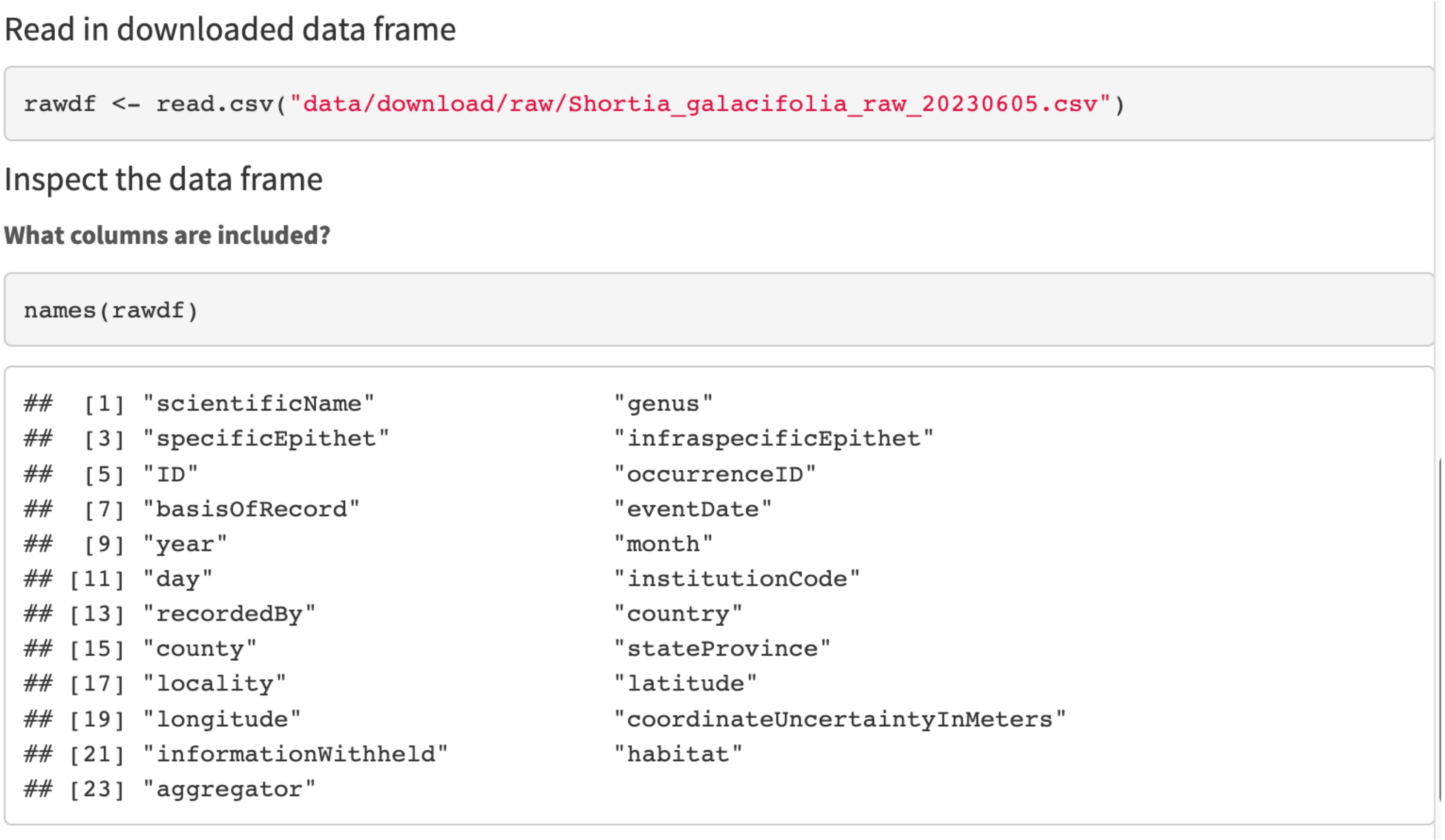
```r
gators_download(synonyms.list = Shortia_galacifolia,
                write.file = TRUE,
                filename = "data/download/raw/Shortia_galacifolia_raw_20230605.csv")
gators_download(synonyms.list = Galax_urceolata,
               write.file = TRUE,
               filename = "data/download/raw/Galax_urceolata_raw_20230605.csv")
gators_download(synonyms.list = Pyxidanthera_barbulata,
               write.file = TRUE,
               filename = "data/download/raw/Pyxidanthera_barbulata_raw_20230605.csv")
gators_download(synonyms.list = Pyxidanthera_brevifolia,
               write.file = TRUE,
               filename = "data/download/raw/Pyxidanthera_brevifolia_raw_20230605.csv")
```

Synonym list

Save csv file

gatoRs

# Quick-look at downloaded files



## Read in downloaded data frame

```
rawdf <- read.csv("data/download/raw/Shortia_galacifolia_raw_20230605.csv")
```

## Inspect the data frame

**What columns are included?**

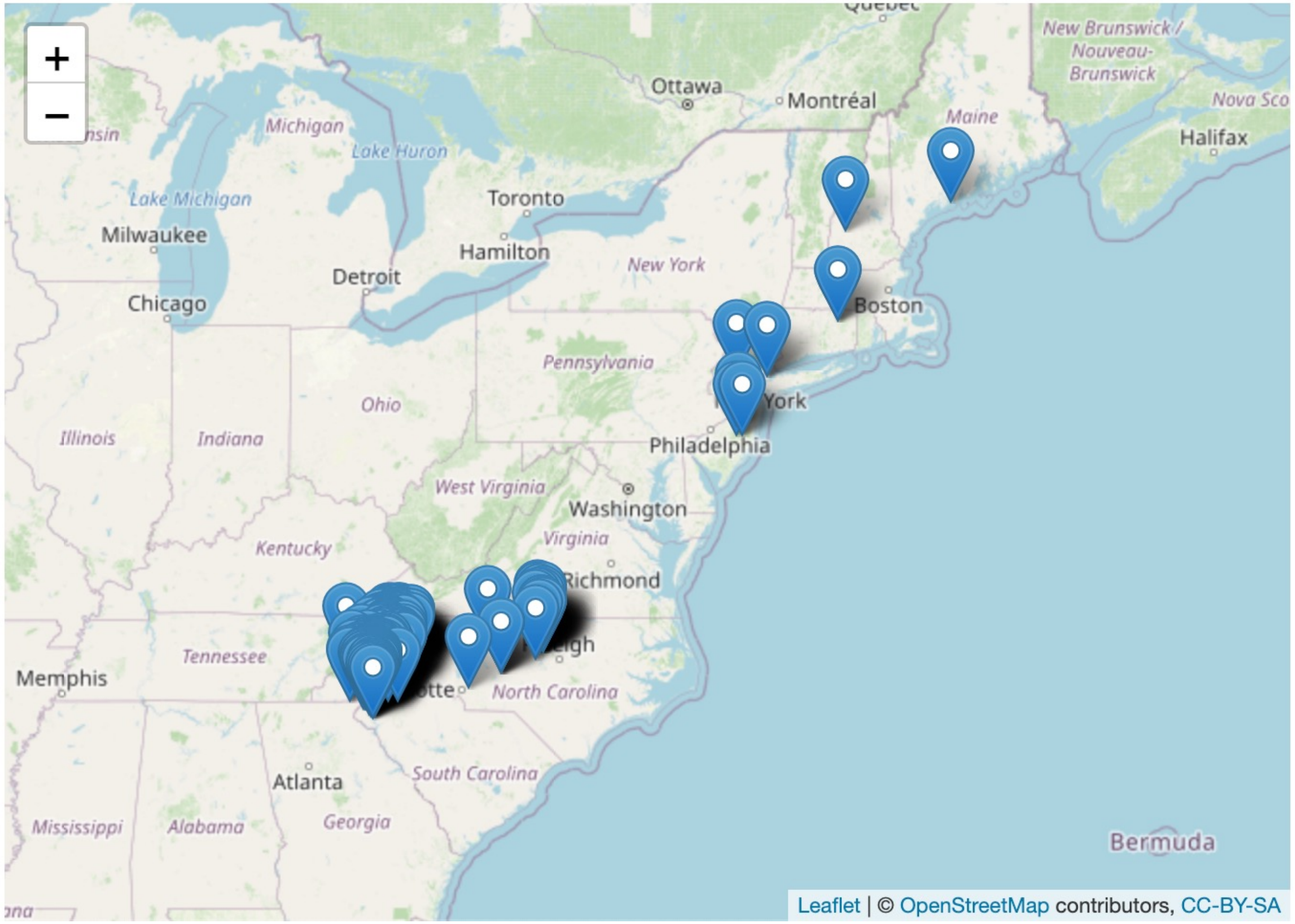
```
names(rawdf)
```

```
##  [1] "scientificName"                "genus"
##  [3] "specificEpithet"               "infraspecificEpithet"
##  [5] "ID"                            "occurrenceID"
##  [7] "basisOfRecord"                 "eventDate"
##  [9] "year"                          "month"
## [11] "day"                           "institutionCode"
## [13] "recordedBy"                    "country"
## [15] "county"                        "stateProvince"
## [17] "locality"                      "latitude"
## [19] "longitude"                     "coordinateUncertaintyInMeters"
## [21] "informationWithheld"           "habitat"
## [23] "aggregator"
```

### Where are these points?

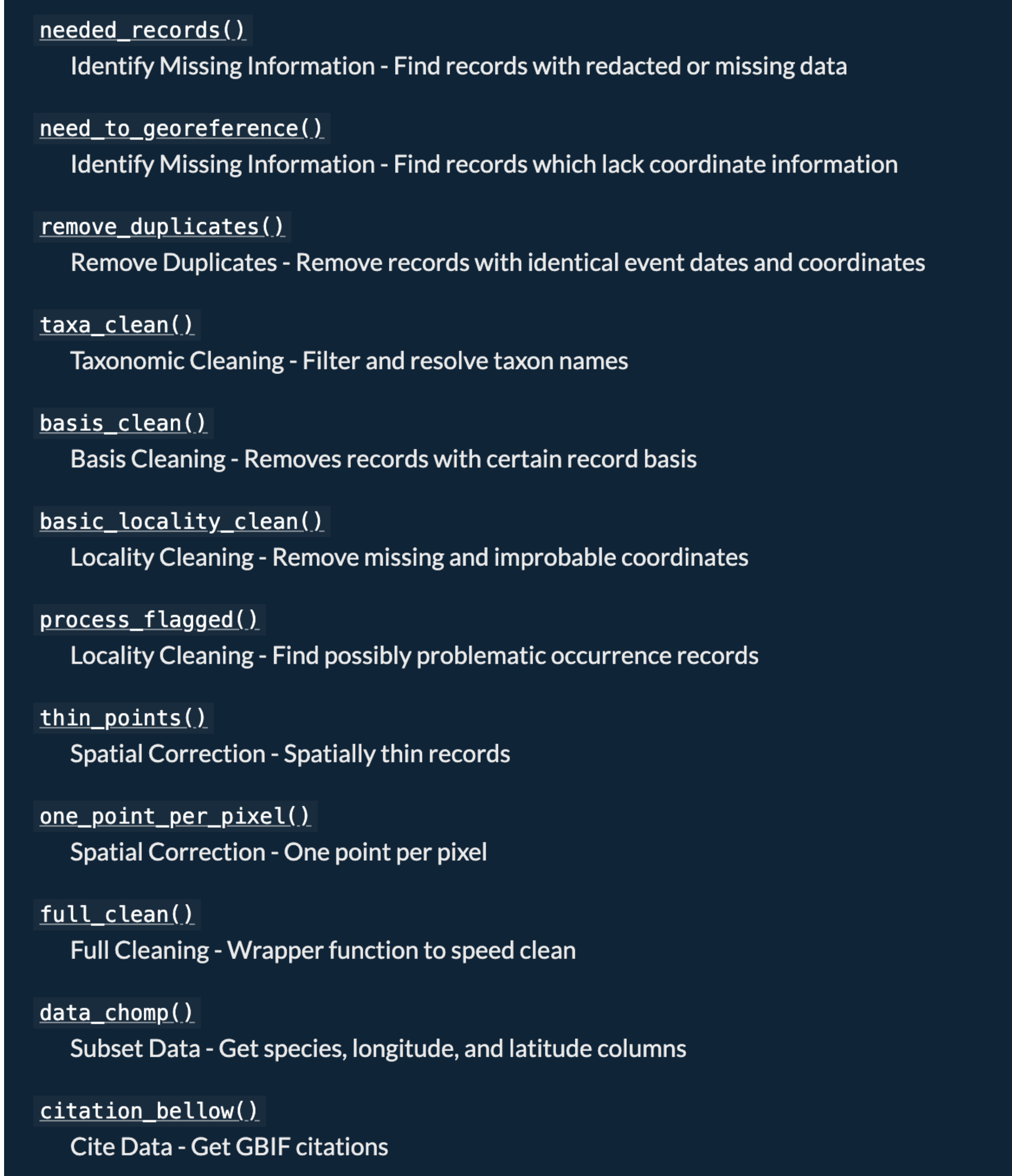
The error message here indicates many points do not have long/lat values (more in 02).

```
leaflet(rawdf) %>%
  addMarkers(label = paste0(rawdf$longitude, ", ", rawdf$latitude)) %>%
  addTiles()
```



Leaflet | © OpenStreetMap contributors, CC-BY-SA gatoRs

Patten et al. *In review*. gatoRs: Geographic and Taxonomic Occurrence R-Based Scrubbing. nataliepatten/gatoRs



`needed_records()`

Identify Missing Information - Find records with redacted or missing data

`need_to_georeference()`

Identify Missing Information - Find records which lack coordinate information

`remove_duplicates()`

Remove Duplicates - Remove records with identical event dates and coordinates

`taxa_clean()`

Taxonomic Cleaning - Filter and resolve taxon names

`basis_clean()`

Basis Cleaning - Removes records with certain record basis

`basic_locality_clean()`

Locality Cleaning - Remove missing and improbable coordinates

`process_flagged()`

Locality Cleaning - Find possibly problematic occurrence records

`thin_points()`

Spatial Correction - Spatially thin records

`one_point_per_pixel()`

Spatial Correction - One point per pixel

`full_clean()`

Full Cleaning - Wrapper function to speed clean

`data_chomp()`

Subset Data - Get species, longitude, and latitude columns

`citation_bellow()`

Cite Data - Get GBIF citations