Webinar Series

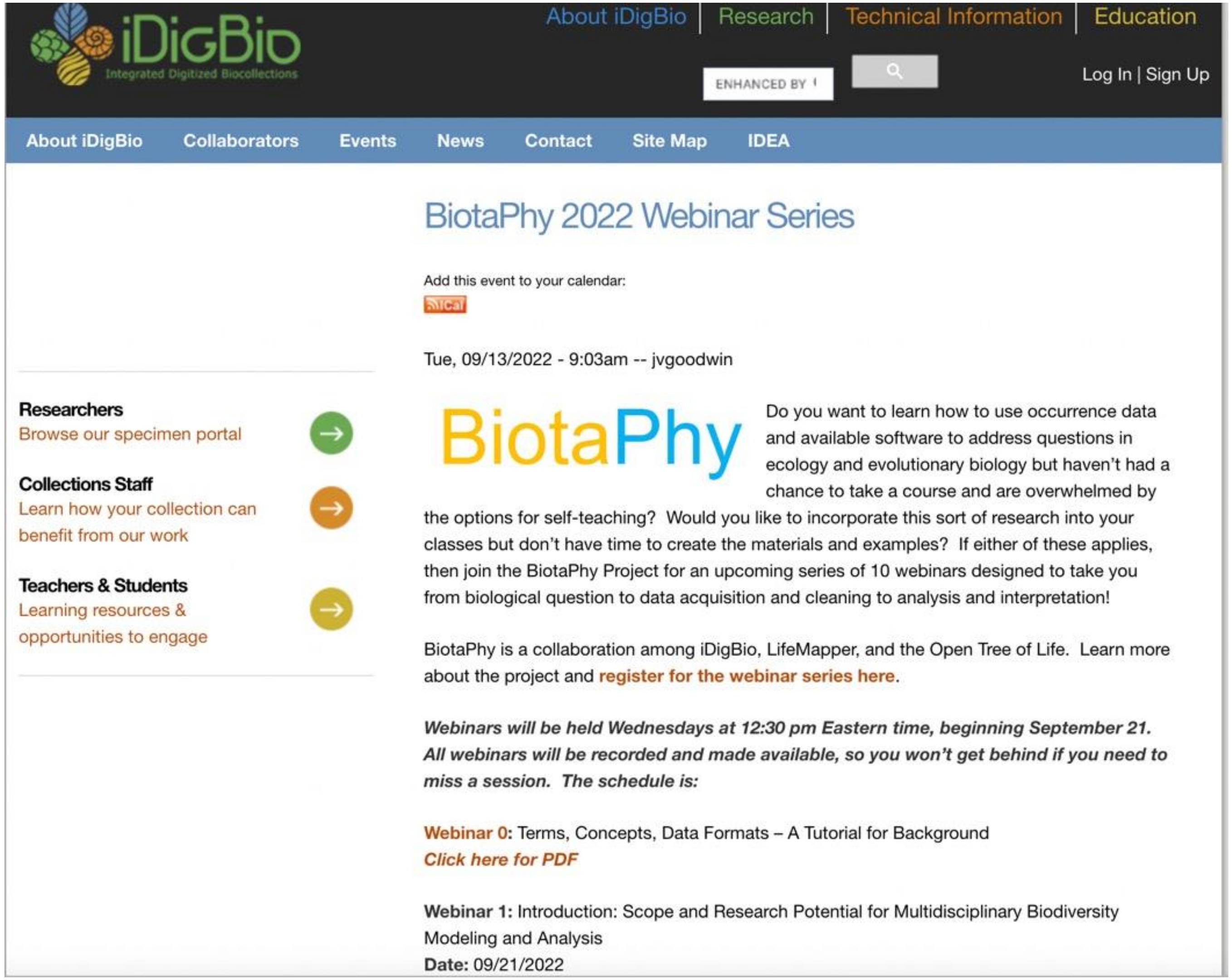
Data Use Skills
Featuring Data from Natural History Collections

September 21-November 30, 2022

https://www.idigbio.org/content/biotaphy-2022-webinar-series

iDigBio:

https://www.idigbio.org/content/biotaphy-2022-webinar-series



BiotaPhy 2022 Webinar Series

Add this event to your calendar:

Tue, 09/13/2022 - 9:03am -- jvgoodwin

Do you want to learn how to use occurrence data and available software to address questions in ecology and evolutionary biology but haven't had a chance to take a course and are overwhelmed by the options for self-teaching? Would you like to incorporate this sort of research into your classes but don't have time to create the materials and examples? If either of these applies, then join the BiotaPhy Project for an upcoming series of 10 webinars designed to take you from biological question to data acquisition and cleaning to analysis and interpretation!

BiotaPhy is a collaboration among iDigBio, LifeMapper, and the Open Tree of Life. Learn more about the project and **register for the webinar series here**.

Webinars will be held Wednesdays at 12:30 pm Eastern time, beginning September 21. All webinars will be recorded and made available, so you won't get behind if you need to miss a session. The schedule is:

Webinar 0: Terms, Concepts, Data Formats – A Tutorial for Background
Click here for PDF

Webinar 1: Introduction: Scope and Research Potential for Multidisciplinary Biodiversity Modeling and Analysis
Date: 09/21/2022

iDigBio.org

Thank You

Maria Cortez
Aimee Stewart

Jill Goodwin
Gil Nelson

Webinar 5

Big Data Munging: Finding, Acquiring, and Preparing Species Occurrence Data and Tree Data

Goals

Learn how to find, manipulate, combine, and use occurrence and tree data in biodiversity analyses

Learning Objectives

Biological Objectives:

✓ **Introduce different data sources (iDigBio, GBIF, SpeciesLink, RainBio, OpenTree, etc.) and data types (occurrence and tree)**

✓ **Showcase options for downloading data directly from portals**

Learning Objectives

Technological Objectives:

✓ **Group retrieved data by taxonomy or other field type.**

✓ **Learn how to combine records from heterogeneous sources.**

Webinar organization

1. Exploring Concepts: why do we need occurrence and tree data, and where can we find them?
2. Demonstrations: how to download occurrence and tree data directly from portals.
3. Exercises: practice automated ways of downloading and treating occurrence and tree data.
4. Session Summary, Q&A, and Discussion

Exploring Concepts

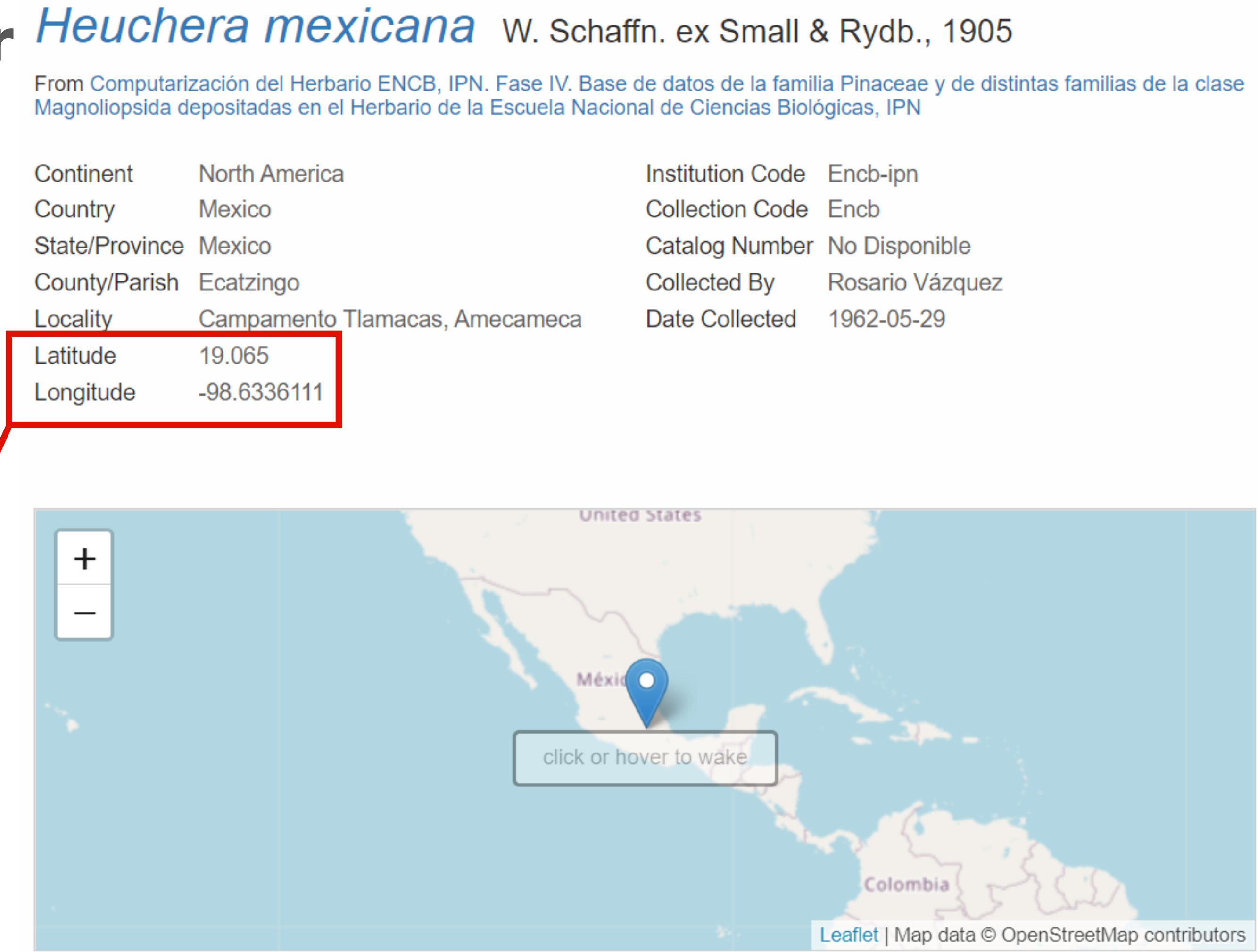
Why do we need occurrence and tree data and where can we find them?

Exploring Concepts

Occurrence data are essential for producing species distribution models, estimating phylogenetic diversity, and more!

Lat: -19.065

Lon: -98.6336111



Heuchera mexicana W. Schaffn. ex Small & Rydb., 1905

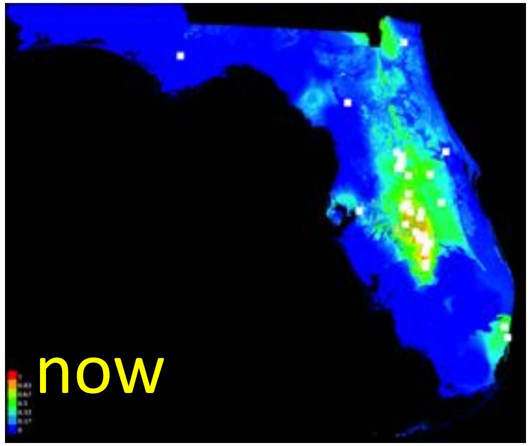
From Computarización del Herbario ENCB, IPN. Fase IV. Base de datos de la familia Pinaceae y de distintas familias de la clase Magnoliopsida depositadas en el Herbario de la Escuela Nacional de Ciencias Biológicas, IPN

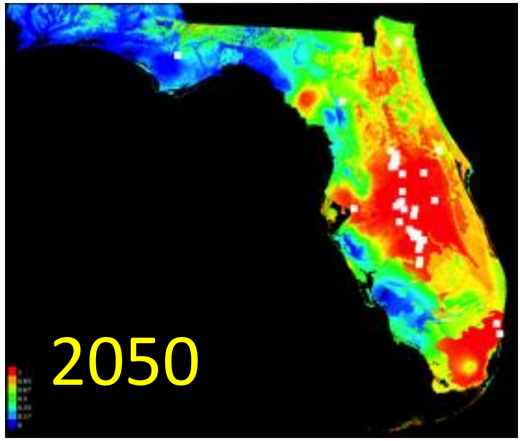
Continent	North America	Institution Code	Encb-ipn
Country	Mexico	Collection Code	Encb
State/Province	Mexico	Catalog Number	No Disponible
County/Parish	Ecatzingo	Collected By	Rosario Vázquez
Locality	Campamento Tlamacas, Amecameca	Date Collected	1962-05-29
Latitude	19.065		
Longitude	-98.6336111		

Exploring Concepts

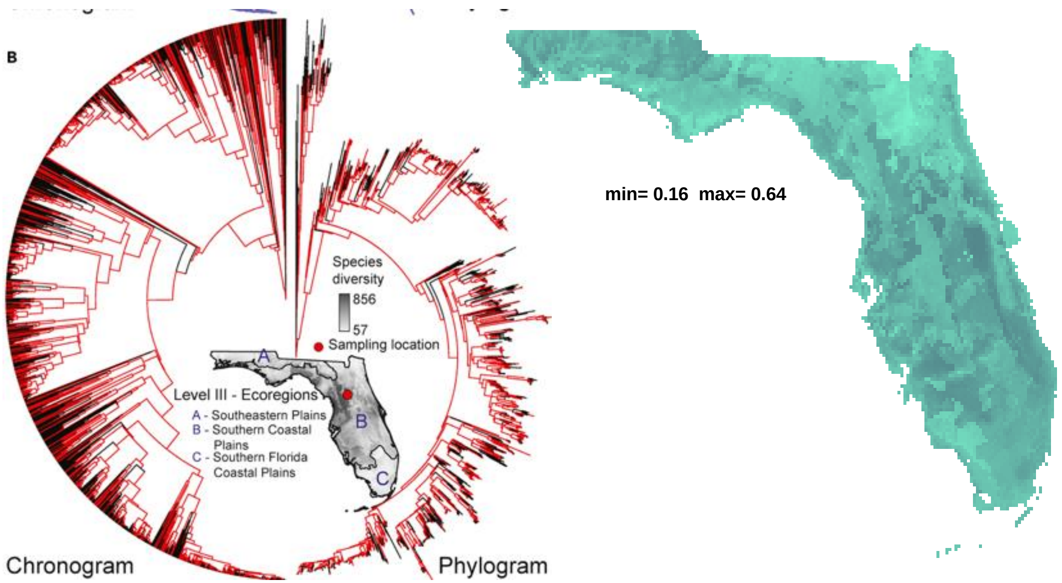
Species distribution models

Prunus geniculata

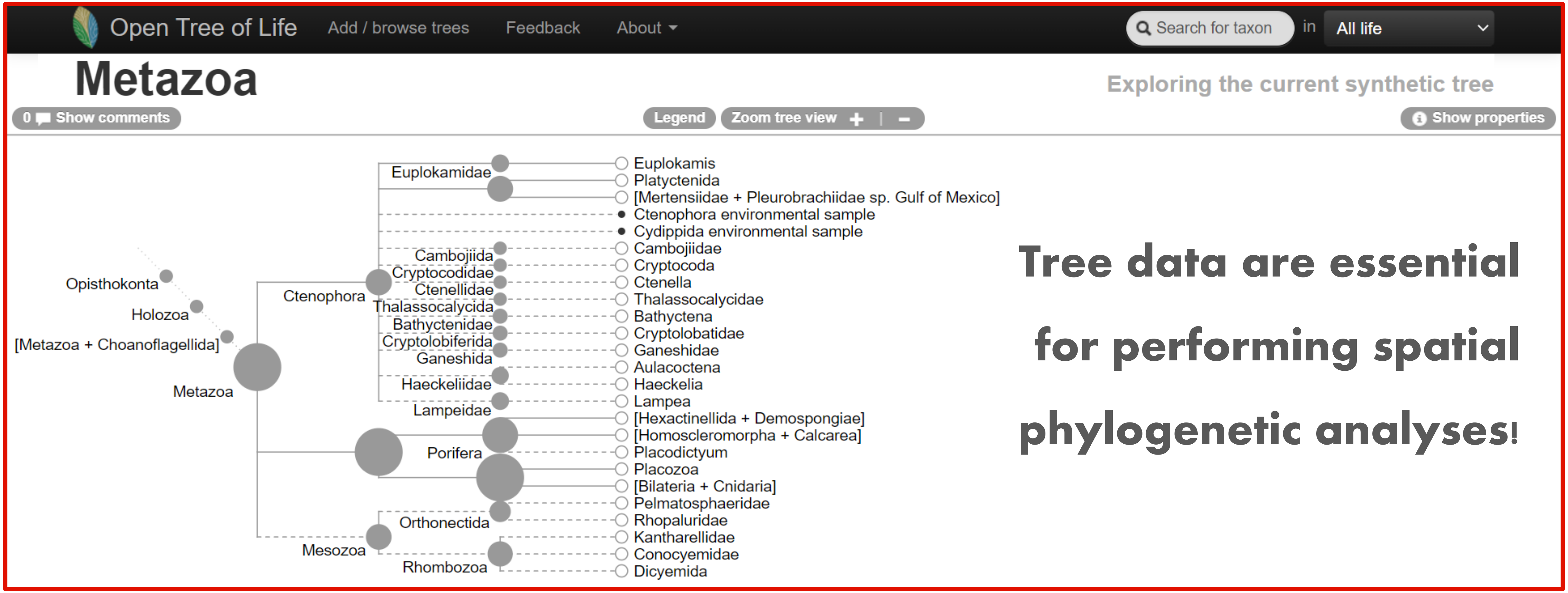




Phylogenetic diversity



Exploring Concepts

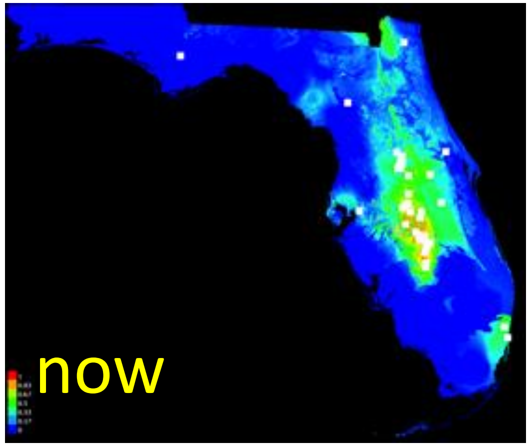


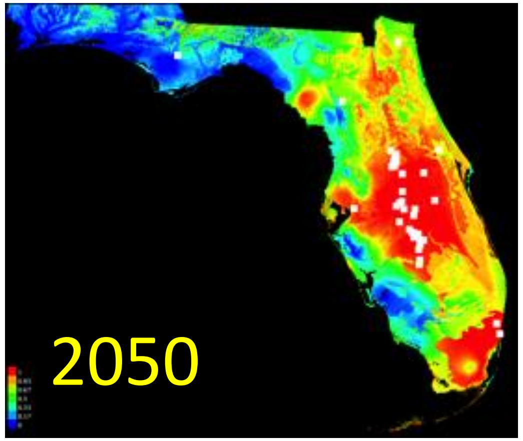
Tree data are essential for performing spatial phylogenetic analyses!

Exploring Concepts

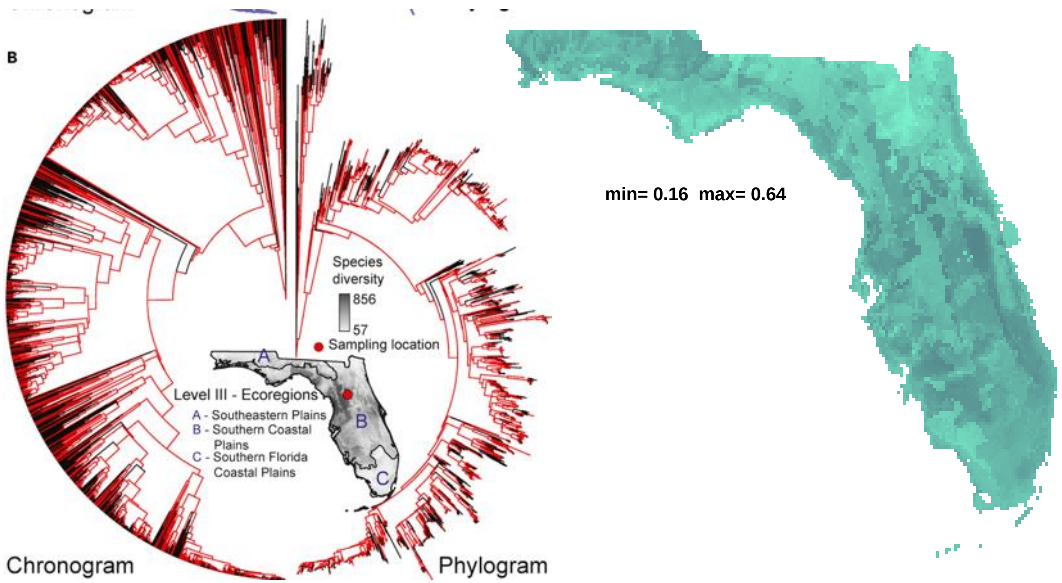
Species distribution models

Prunus geniculata





Phylogenetic diversity

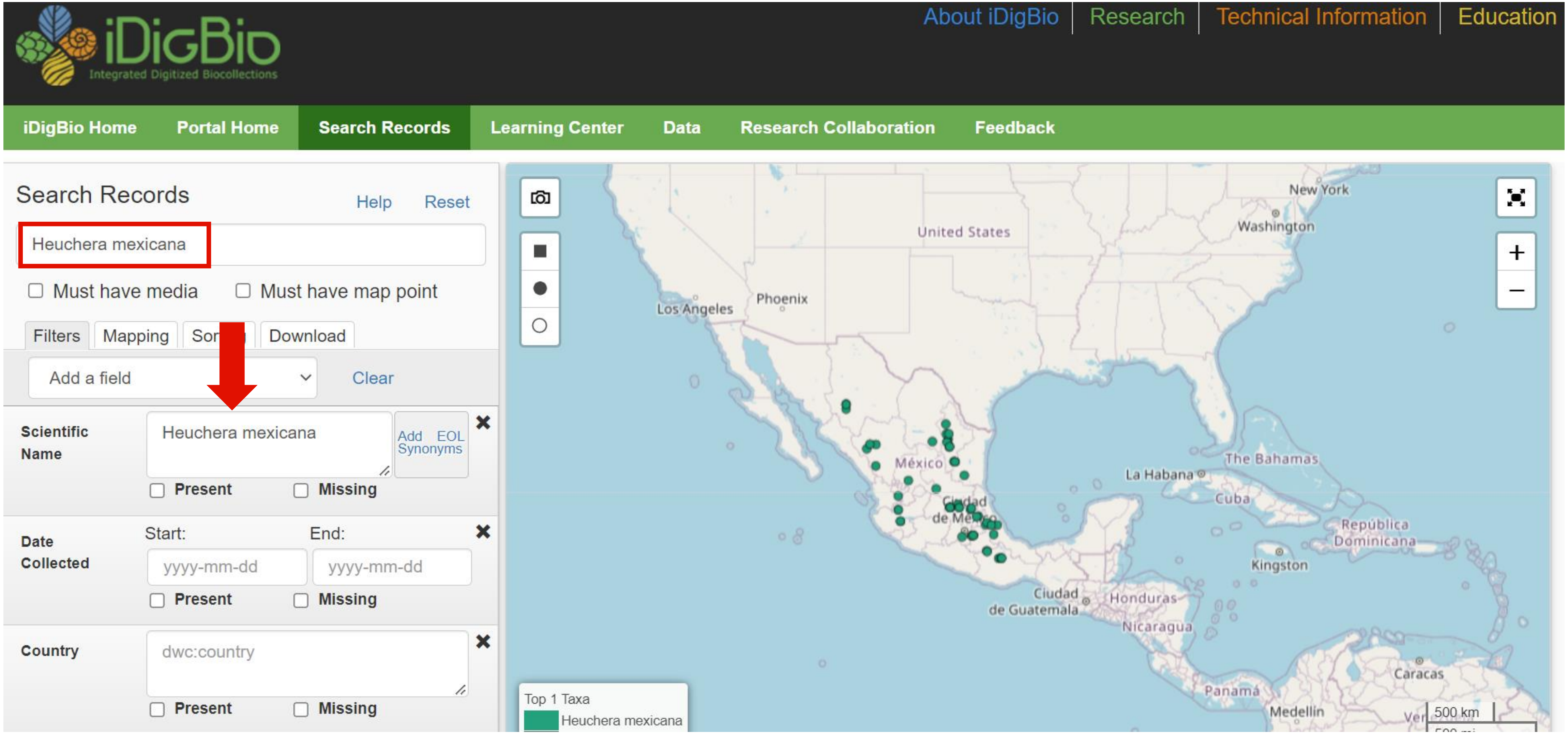


Demonstration: Downloading data from portals

Demonstration: how to download data directly from portals!

- Useful for smaller datasets
- Useful for demonstrations
- Useful as a teaching tool
- Easy to visualize the data

Demonstration: Downloading data from portals



Search Records Help Reset

Heuchera mexicana

☐ Must have media ☐ Must have map point

Filters | Mapping | Sorting | Download

Add a field Clear

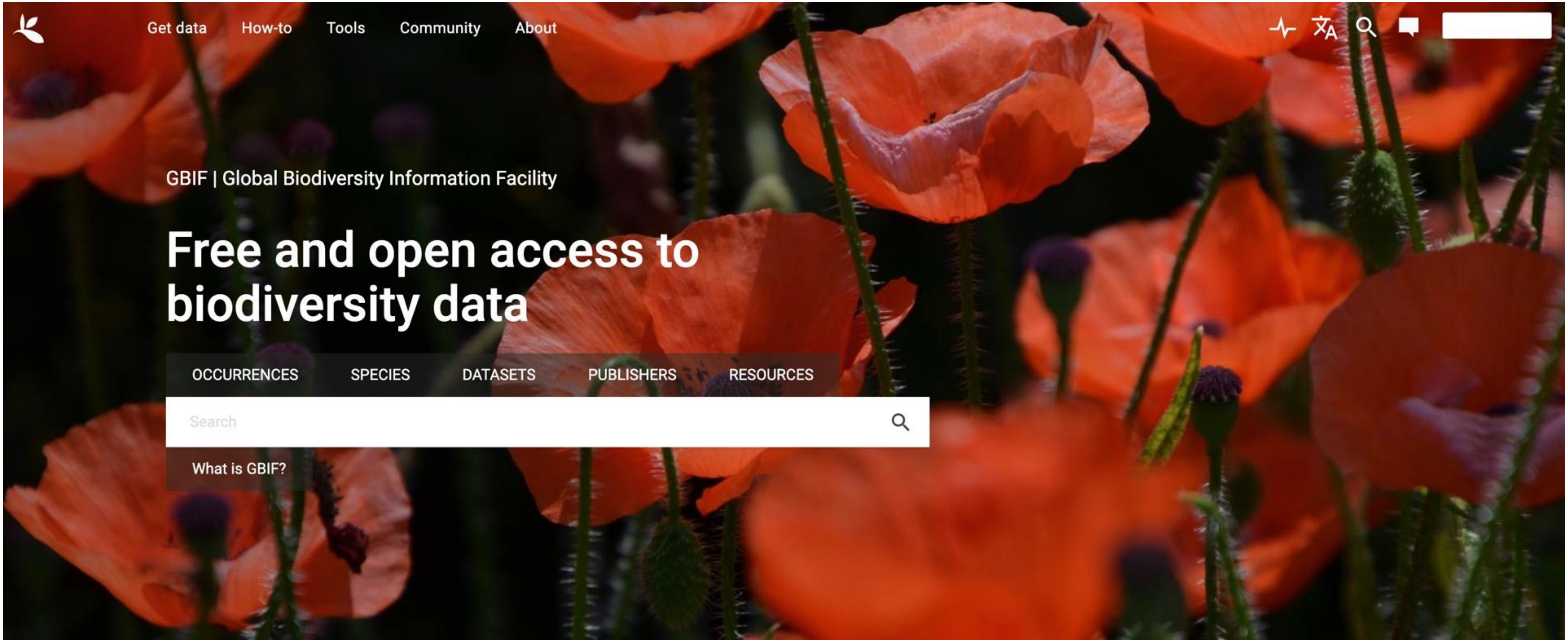
Scientific Name: Heuchera mexicana (Add EOL Synonyms) ✖
☐ Present ☐ Missing

Date Collected: Start: yyyy-mm-dd End: yyyy-mm-dd ✖
☐ Present ☐ Missing

Country: dwc:country ✖
☐ Present ☐ Missing

Top 1 Taxa
🟩 Heuchera mexicana

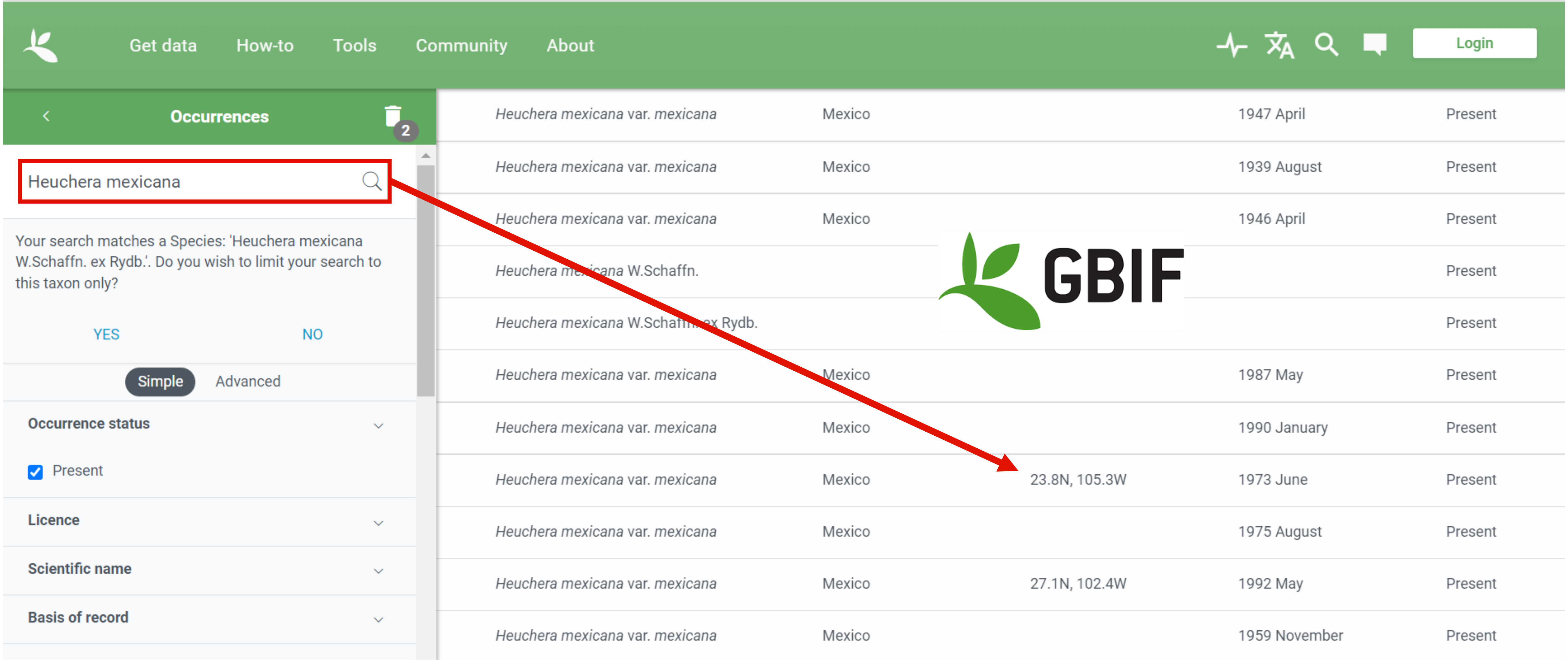
Demonstration: Downloading data from portals



GBIF | Global Biodiversity Information Facility

Free and open access to biodiversity data

Demonstration: Downloading data from portals



Heuchera mexicana

Your search matches a Species: 'Heuchera mexicana W.Schaffn. ex Rydb.'. Do you wish to limit your search to this taxon only?

YES NO

Simple Advanced

Occurrence status

☑ Present

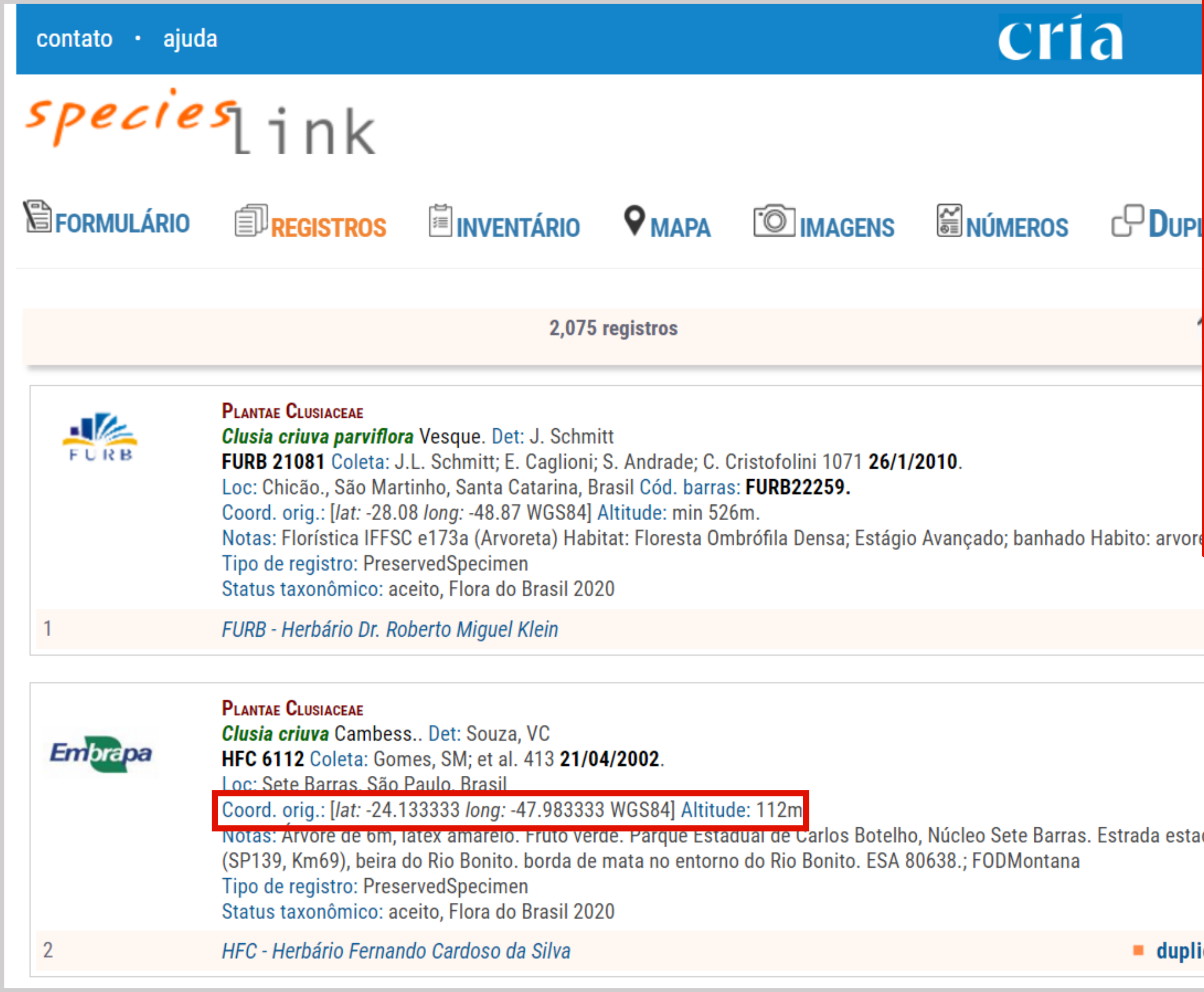
Licence

Scientific name

Basis of record

Heuchera mexicana var. *mexicana*	Mexico		1947 April	Present
Heuchera mexicana var. *mexicana*	Mexico		1939 August	Present
Heuchera mexicana var. *mexicana*	Mexico		1946 April	Present
Heuchera mexicana W.Schaffn.				Present
Heuchera mexicana W.Schaffn. ex Rydb.				Present
Heuchera mexicana var. *mexicana*	Mexico		1987 May	Present
Heuchera mexicana var. *mexicana*	Mexico		1990 January	Present
Heuchera mexicana var. *mexicana*	Mexico	23.8N, 105.3W	1973 June	Present
Heuchera mexicana var. *mexicana*	Mexico		1975 August	Present
Heuchera mexicana var. *mexicana*	Mexico	27.1N, 102.4W	1992 May	Present
Heuchera mexicana var. *mexicana*	Mexico		1959 November	Present

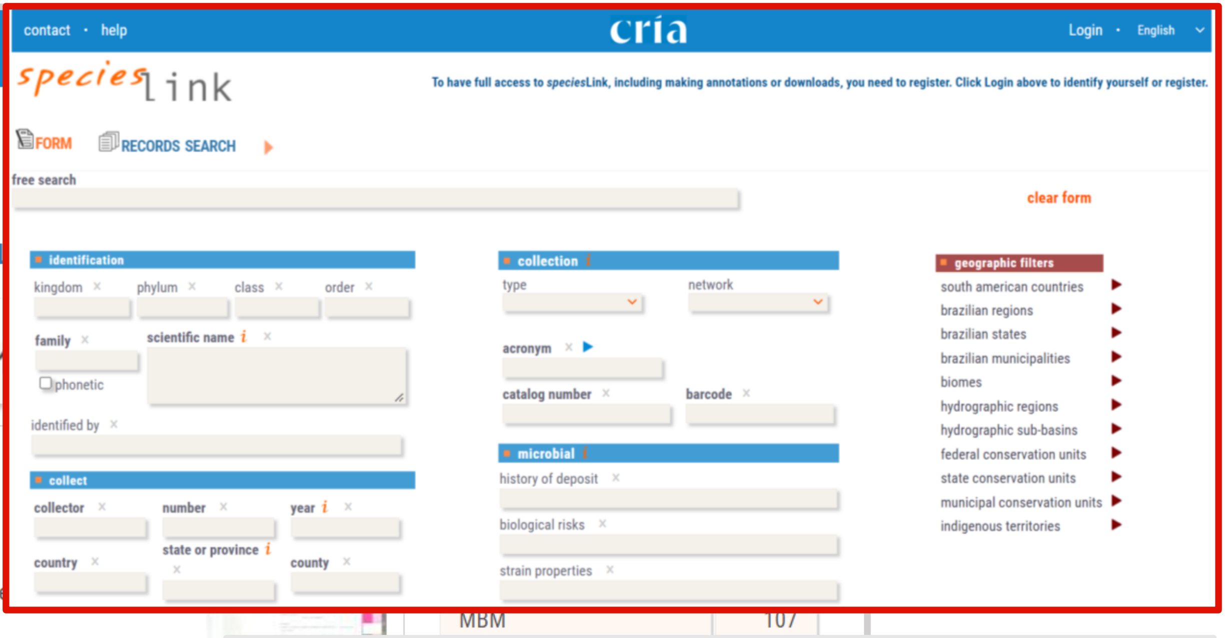
Demonstration: Downloading data from portals



2,075 registros

Plantae Clusiaceae
Clusia criuva parviflora Vesque. Det: J. Schmitt
FURB 21081 Coleta: J.L. Schmitt; E. Caglioni; S. Andrade; C. Cristofolini 1071 **26/1/2010**.
Loc: Chicão., São Martinho, Santa Catarina, Brasil Cód. barras: **FURB22259.**
Coord. orig.: [*lat:* -28.08 *long:* -48.87 WGS84] Altitude: min 526m.
Notas: Florística IFFSC e173a (Arvoreta) Habitat: Floresta Ombrófila Densa; Estágio Avançado; banhado Habito: arvore
Tipo de registro: PreservedSpecimen
Status taxonômico: aceito, Flora do Brasil 2020
1 *FURB - Herbário Dr. Roberto Miguel Klein* ■ nova anotaç

Plantae Clusiaceae
Clusia criuva Cambess.. Det: Souza, VC
HFC 6112 Coleta: Gomes, SM; et al. 413 **21/04/2002**.
Loc: Sete Barras, São Paulo, Brasil
Coord. orig.: [*lat:* -24.133333 *long:* -47.983333 WGS84] Altitude: 112m
Notas: Árvore de 6m, látex amarelo. Fruto verde. Parque Estadual de Carlos Botelho, Núcleo Sete Barras. Estrada estadual São Miguel Arcanjo (SP139, Km69), beira do Rio Bonito. borda de mata no entorno do Rio Bonito. ESA 80638.; FODMontana
Tipo de registro: PreservedSpecimen
Status taxonômico: aceito, Flora do Brasil 2020
2 *HFC - Herbário Fernando Cardoso da Silva* ■ duplicatas ■ nova anotaç



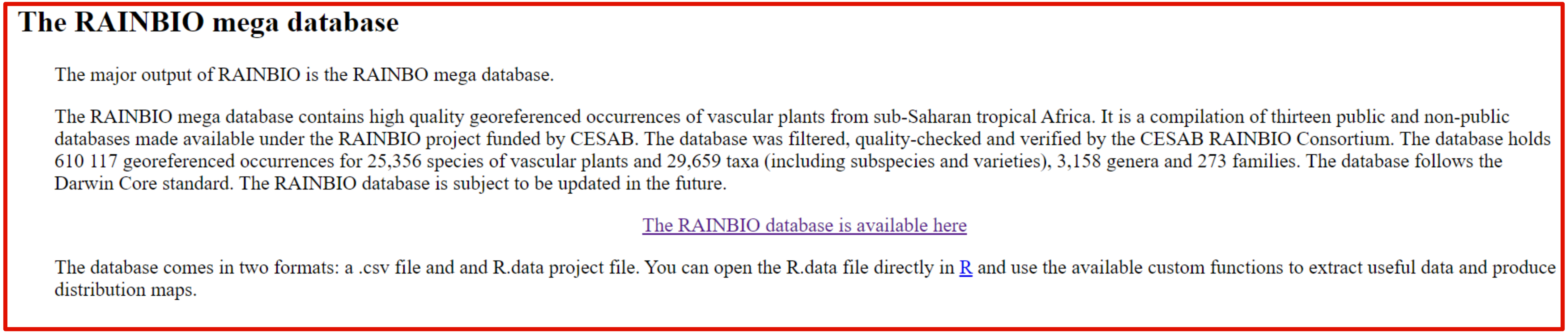
To have full access to *species*Link, including making annotations or downloads, you need to register. Click Login above to identify yourself or register.

> ***"…speciesLink promotes free and open access to data, information, and tools…"***

most databases included are Brazilian

Demonstration: Downloading data from portals

> *"…database contains high quality georeferenced occurrences of vascular plants from sub-Saharan tropical Africa."*



The RAINBIO mega database

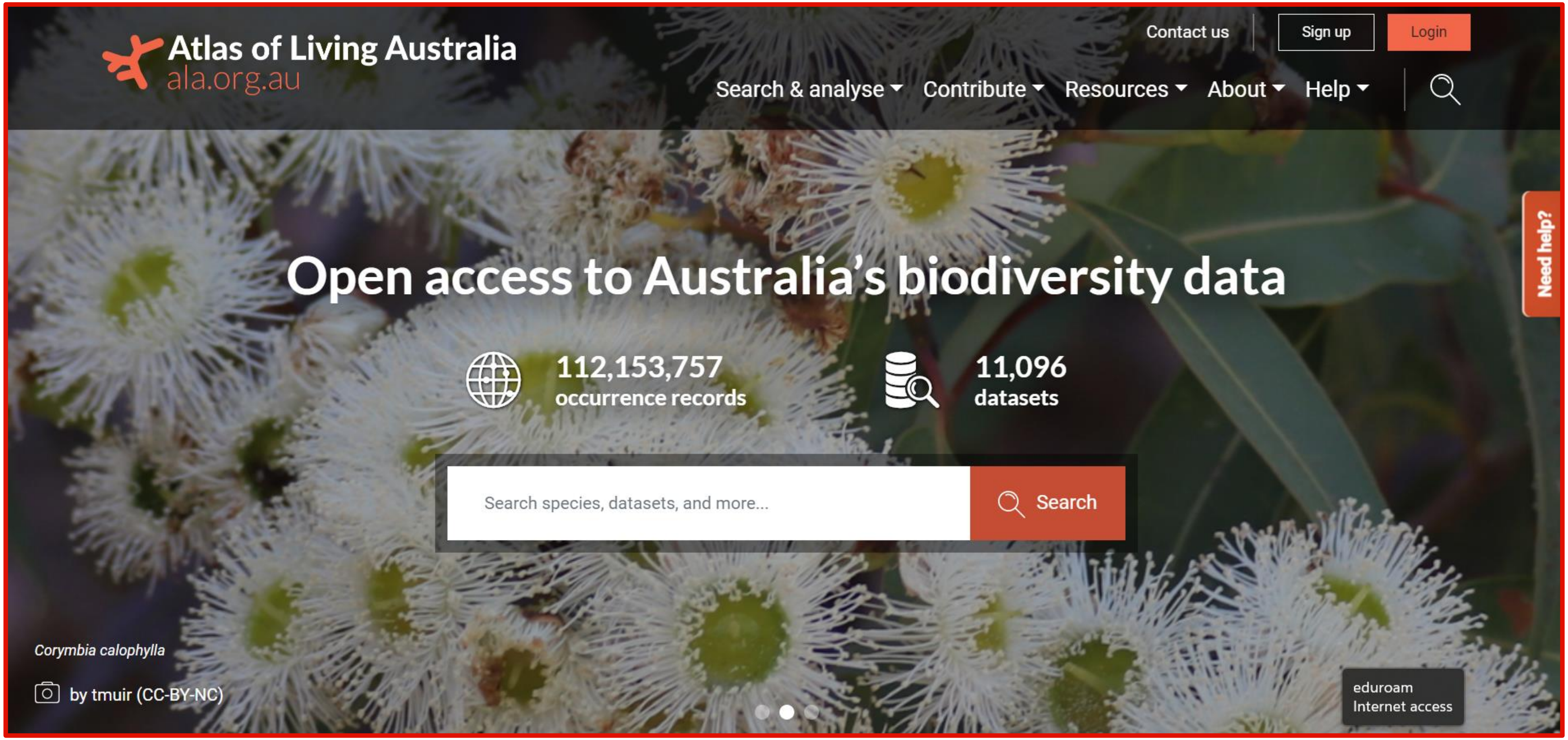
The major output of RAINBIO is the RAINBO mega database.

The RAINBIO mega database contains high quality georeferenced occurrences of vascular plants from sub-Saharan tropical Africa. It is a compilation of thirteen public and non-public databases made available under the RAINBIO project funded by CESAB. The database was filtered, quality-checked and verified by the CESAB RAINBIO Consortium. The database holds 610 117 georeferenced occurrences for 25,356 species of vascular plants and 29,659 taxa (including subspecies and varieties), 3,158 genera and 273 families. The database follows the Darwin Core standard. The RAINBIO database is subject to be updated in the future.

The RAINBIO database is available here

The database comes in two formats: a .csv file and and R.data project file. You can open the R.data file directly in R and use the available custom functions to extract useful data and produce distribution maps.

Demonstration: Downloading data from portals



Atlas of Living Australia
ala.org.au

Open access to Australia's biodiversity data

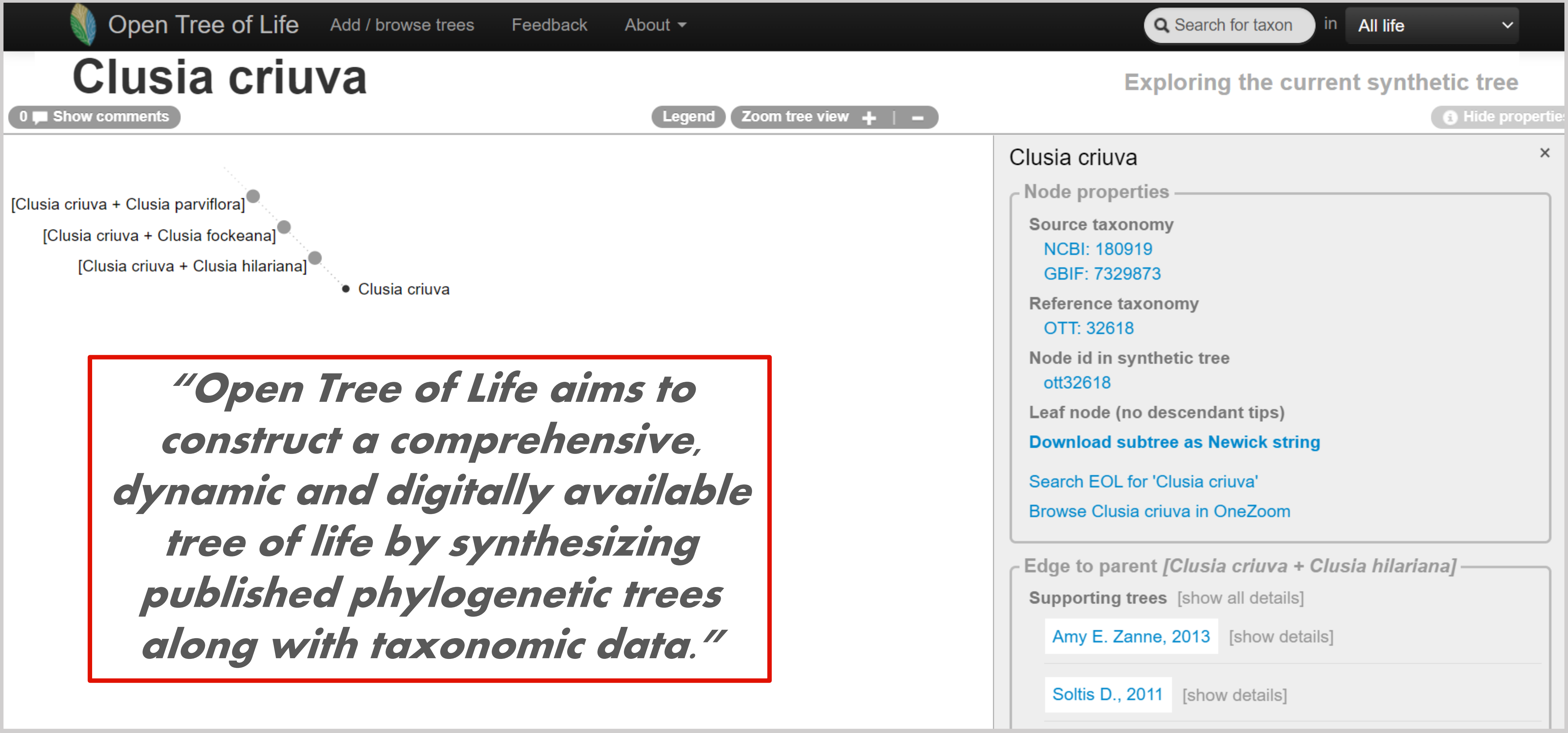
112,153,757 occurrence records

11,096 datasets

Corymbia calophylla

by tmuir (CC-BY-NC)

Demonstration: Downloading data from portals



Clusia criuva

Exploring the current synthetic tree

Clusia criuva

Node properties

Source taxonomy
NCBI: 180919
GBIF: 7329873

Reference taxonomy
OTT: 32618

Node id in synthetic tree
ott32618

Leaf node (no descendant tips)

Download subtree as Newick string

Search EOL for 'Clusia criuva'
Browse Clusia criuva in OneZoom

Edge to parent *[Clusia criuva + Clusia hilariana]*

Supporting trees [show all details]

Amy E. Zanne, 2013 [show details]

Soltis D., 2011 [show details]

> ***"Open Tree of Life aims to construct a comprehensive, dynamic and digitally available tree of life by synthesizing published phylogenetic trees along with taxonomic data."***

Time to Exercise!

What happens when there is a large dataset? Should we download and treat 40,000 records individually?

We use BiotaPhy tools to automate 😅 occurrence and tree data munging!

Time to Exercise!

Let’s put the automated framework developed by BiotaPhy to the test!

How to split and merge occurrence data by species:

3 steps:

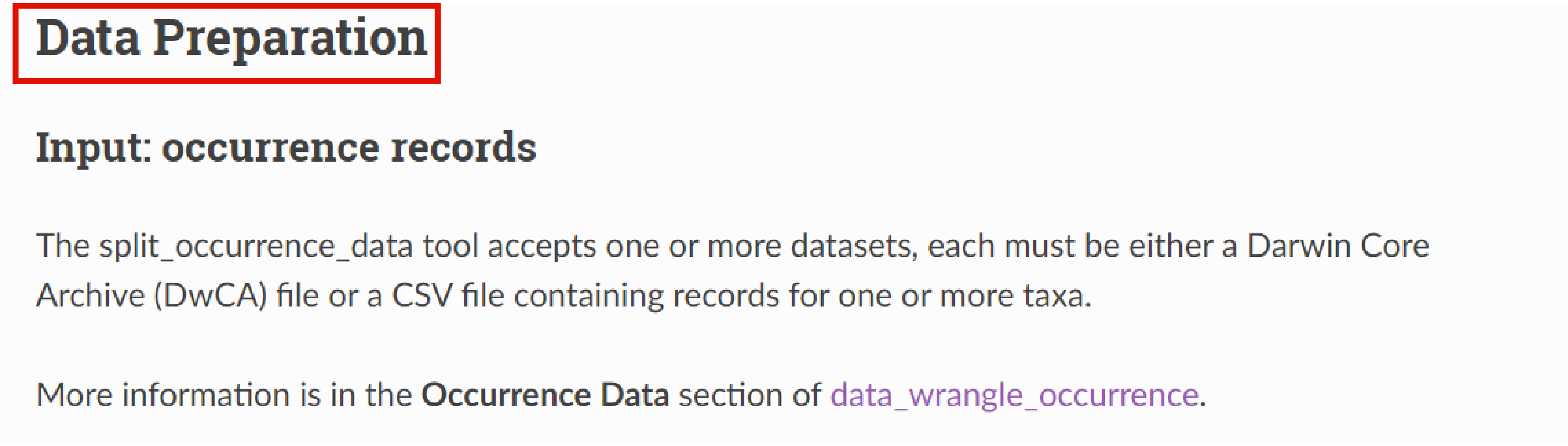
- ✓ Data Preparation
- ✓ Run Tutorial
- ✓ Inspect Output

Input: occurrence records

Input: Wrangler configuration file

Input: Script parameter file

Data Preparation: occurrence records



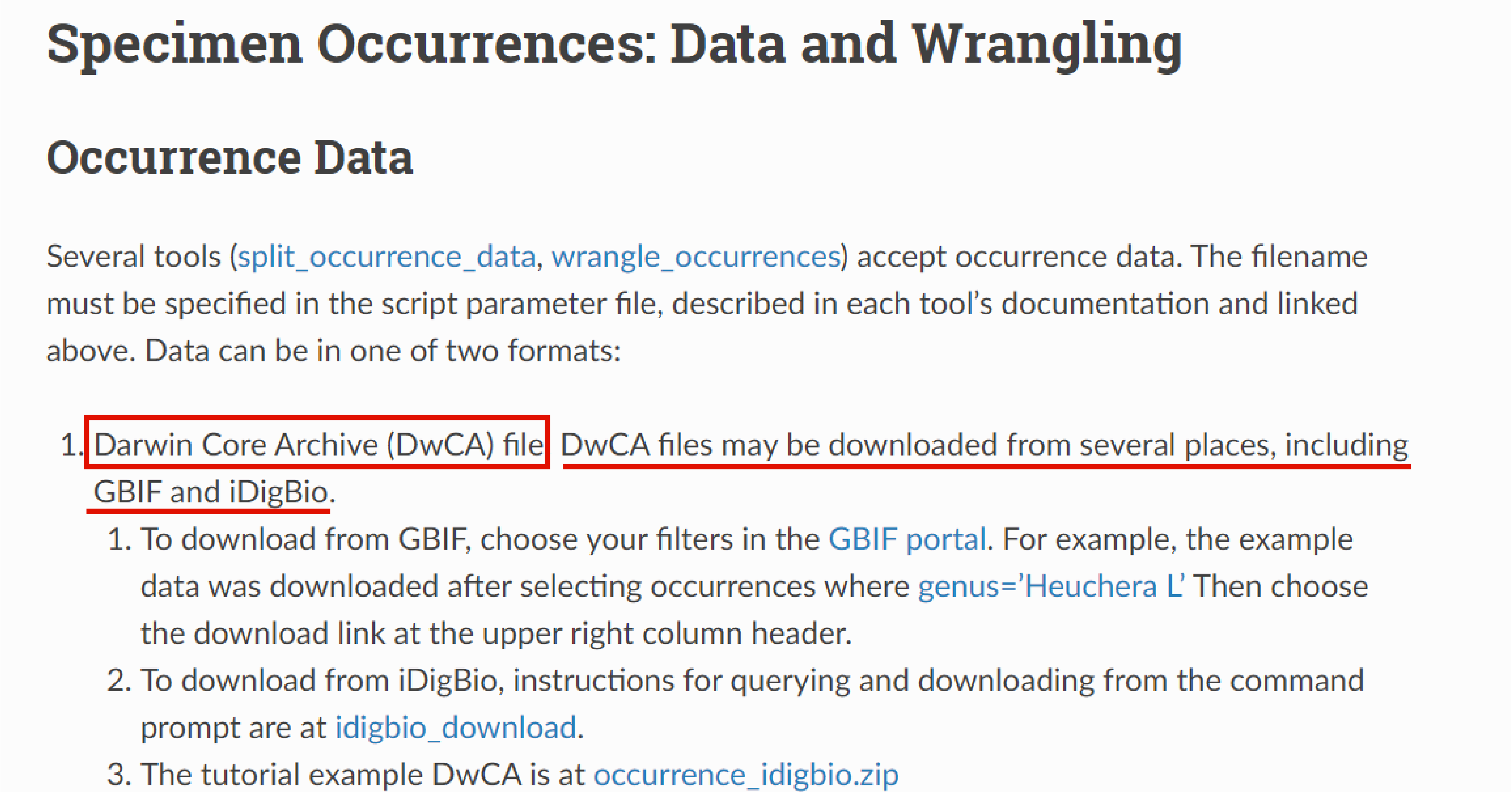
Data Preparation

Input: occurrence records

The split_occurrence_data tool accepts one or more datasets, each must be either a Darwin Core Archive (DwCA) file or a CSV file containing records for one or more taxa.

More information is in the **Occurrence Data** section of data_wrangle_occurrence.

Data Preparation: occurrence records



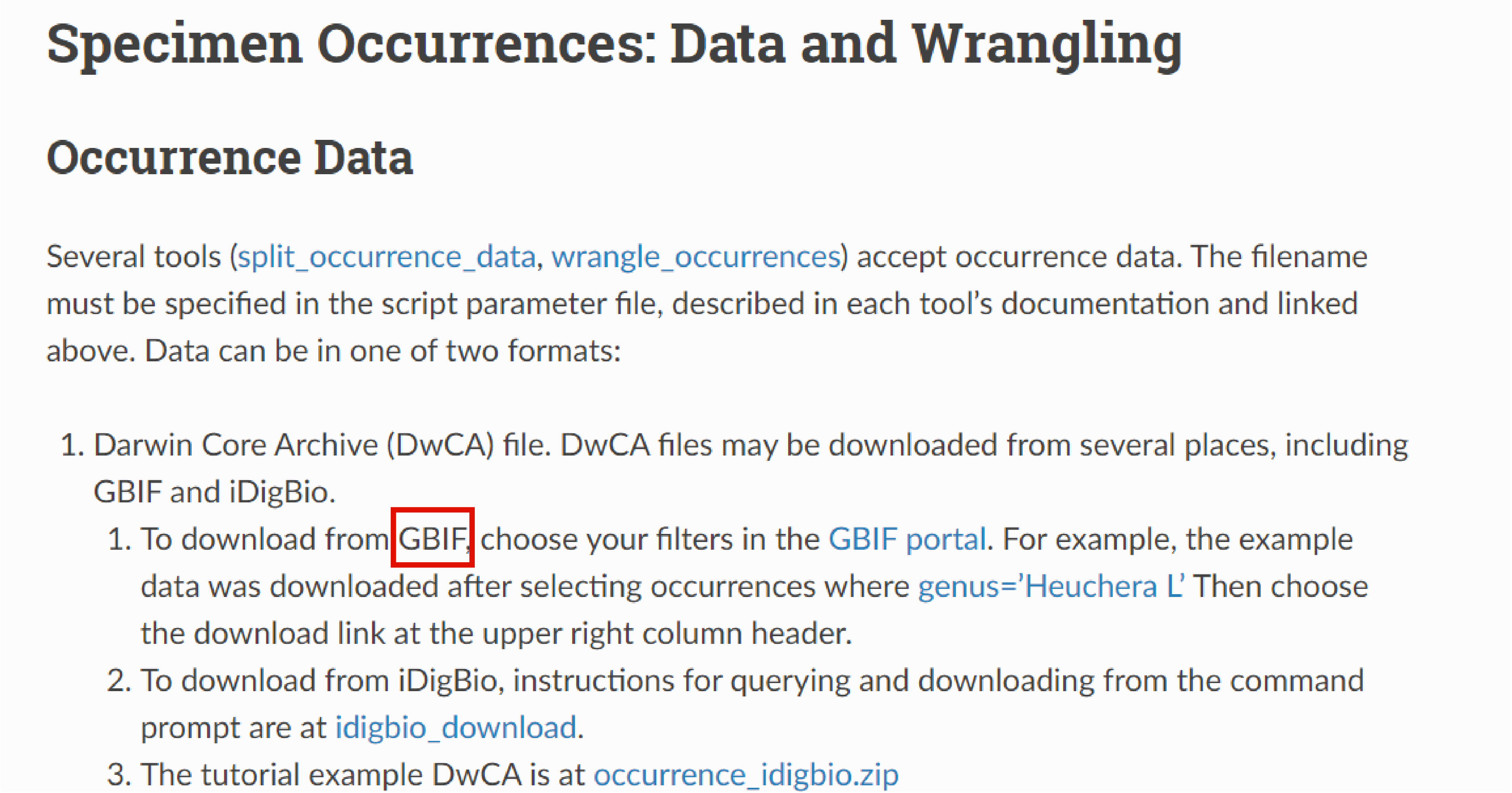
Specimen Occurrences: Data and Wrangling

Occurrence Data

Several tools (split_occurrence_data, wrangle_occurrences) accept occurrence data. The filename must be specified in the script parameter file, described in each tool's documentation and linked above. Data can be in one of two formats:

1. Darwin Core Archive (DwCA) file DwCA files may be downloaded from several places, including GBIF and iDigBio.
    1. To download from GBIF, choose your filters in the GBIF portal. For example, the example data was downloaded after selecting occurrences where genus='Heuchera L' Then choose the download link at the upper right column header.
    2. To download from iDigBio, instructions for querying and downloading from the command prompt are at idigbio_download.
    3. The tutorial example DwCA is at occurrence_idigbio.zip

Data Preparation: occurrence records



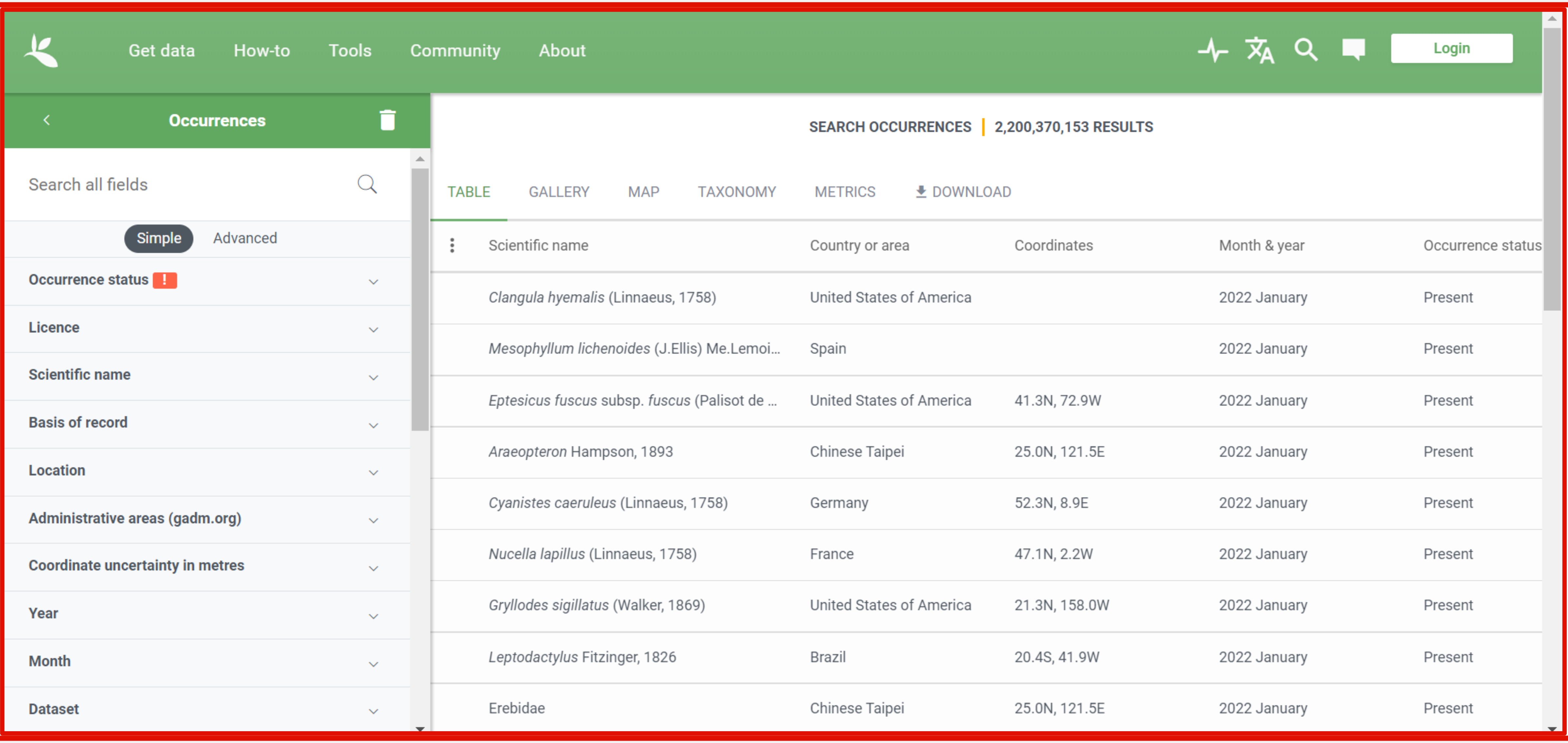
Specimen Occurrences: Data and Wrangling

Occurrence Data

Several tools (split_occurrence_data, wrangle_occurrences) accept occurrence data. The filename must be specified in the script parameter file, described in each tool’s documentation and linked above. Data can be in one of two formats:

1. Darwin Core Archive (DwCA) file. DwCA files may be downloaded from several places, including GBIF and iDigBio.
   1. To download from GBIF, choose your filters in the GBIF portal. For example, the example data was downloaded after selecting occurrences where genus=’Heuchera L’ Then choose the download link at the upper right column header.
   2. To download from iDigBio, instructions for querying and downloading from the command prompt are at idigbio_download.
   3. The tutorial example DwCA is at occurrence_idigbio.zip

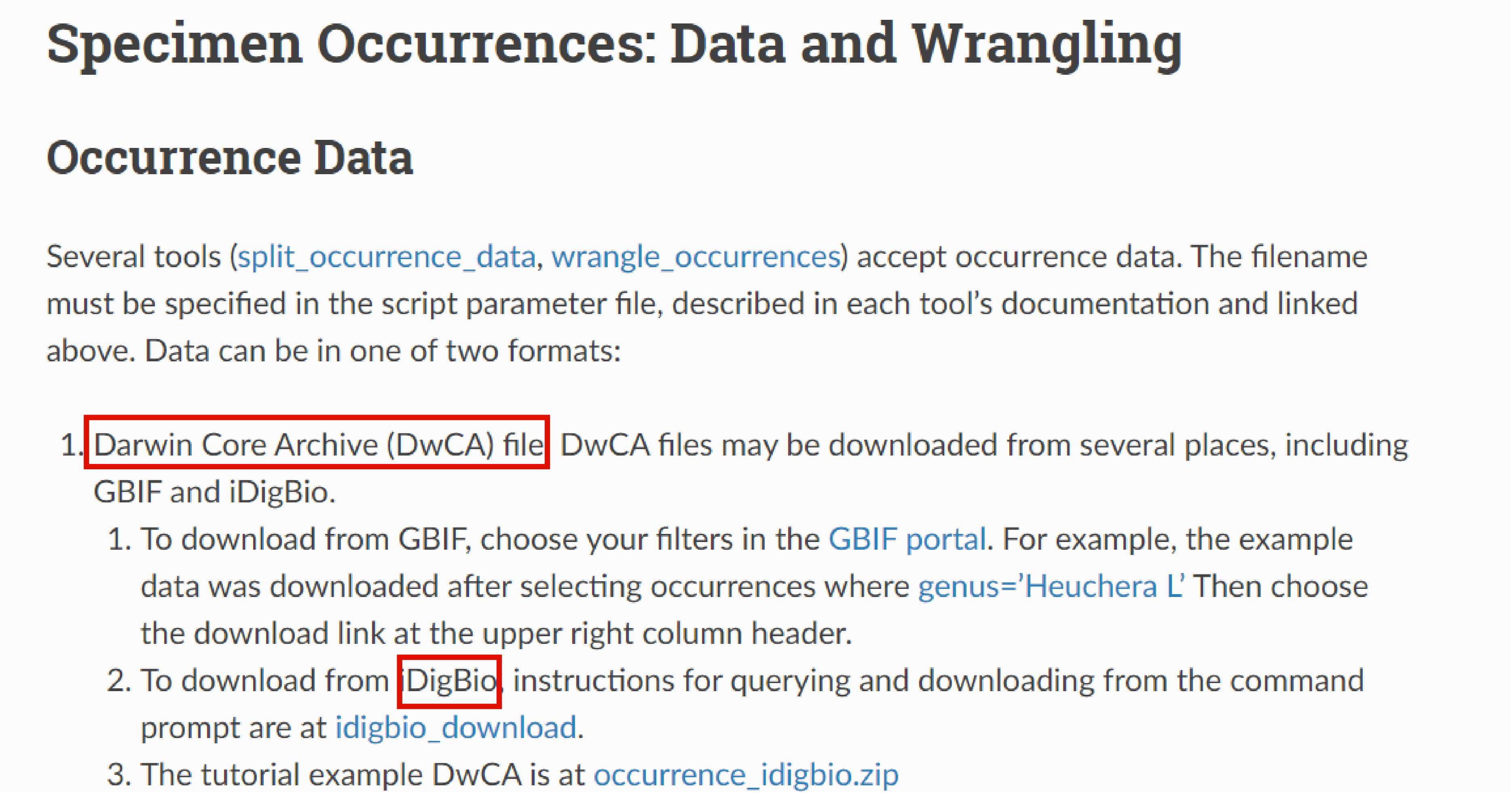
Data Preparation: occurrence records



SEARCH OCCURRENCES | 2,200,370,153 RESULTS

Scientific name	Country or area	Coordinates	Month & year	Occurrence status
Clangula hyemalis (Linnaeus, 1758)	United States of America		2022 January	Present
Mesophyllum lichenoides (J.Ellis) Me.Lemoi...	Spain		2022 January	Present
Eptesicus fuscus subsp. *fuscus* (Palisot de ...	United States of America	41.3N, 72.9W	2022 January	Present
Araeopteron Hampson, 1893	Chinese Taipei	25.0N, 121.5E	2022 January	Present
Cyanistes caeruleus (Linnaeus, 1758)	Germany	52.3N, 8.9E	2022 January	Present
Nucella lapillus (Linnaeus, 1758)	France	47.1N, 2.2W	2022 January	Present
Gryllodes sigillatus (Walker, 1869)	United States of America	21.3N, 158.0W	2022 January	Present
Leptodactylus Fitzinger, 1826	Brazil	20.4S, 41.9W	2022 January	Present
Erebidae	Chinese Taipei	25.0N, 121.5E	2022 January	Present

Data Preparation: occurrence records



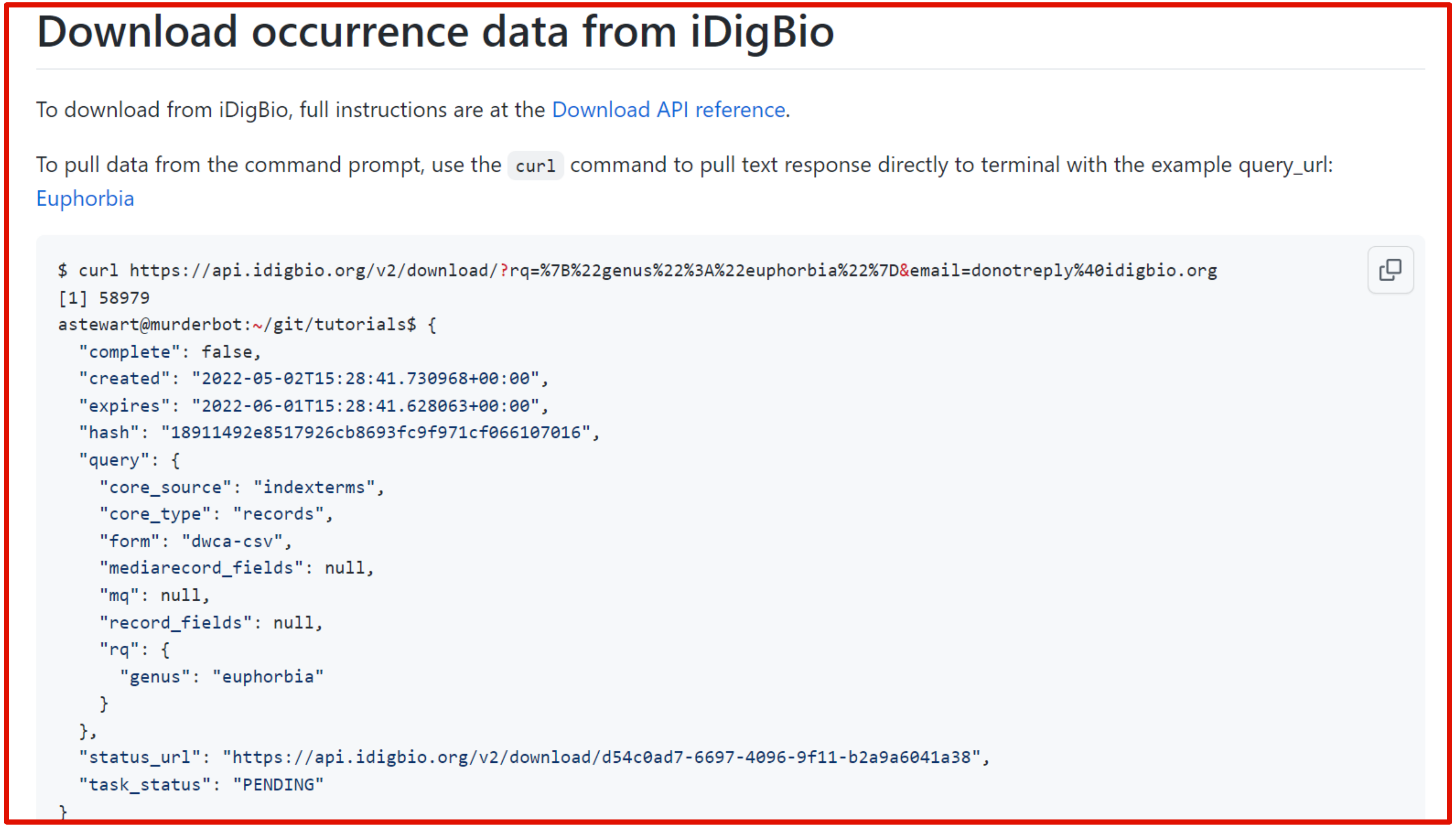
Specimen Occurrences: Data and Wrangling

Occurrence Data

Several tools (split_occurrence_data, wrangle_occurrences) accept occurrence data. The filename must be specified in the script parameter file, described in each tool's documentation and linked above. Data can be in one of two formats:

1. Darwin Core Archive (DwCA) file DwCA files may be downloaded from several places, including GBIF and iDigBio.
    1. To download from GBIF, choose your filters in the GBIF portal. For example, the example data was downloaded after selecting occurrences where genus='Heuchera L' Then choose the download link at the upper right column header.
    2. To download from iDigBio, instructions for querying and downloading from the command prompt are at idigbio_download.
    3. The tutorial example DwCA is at occurrence_idigbio.zip

Data Preparation: occurrence records



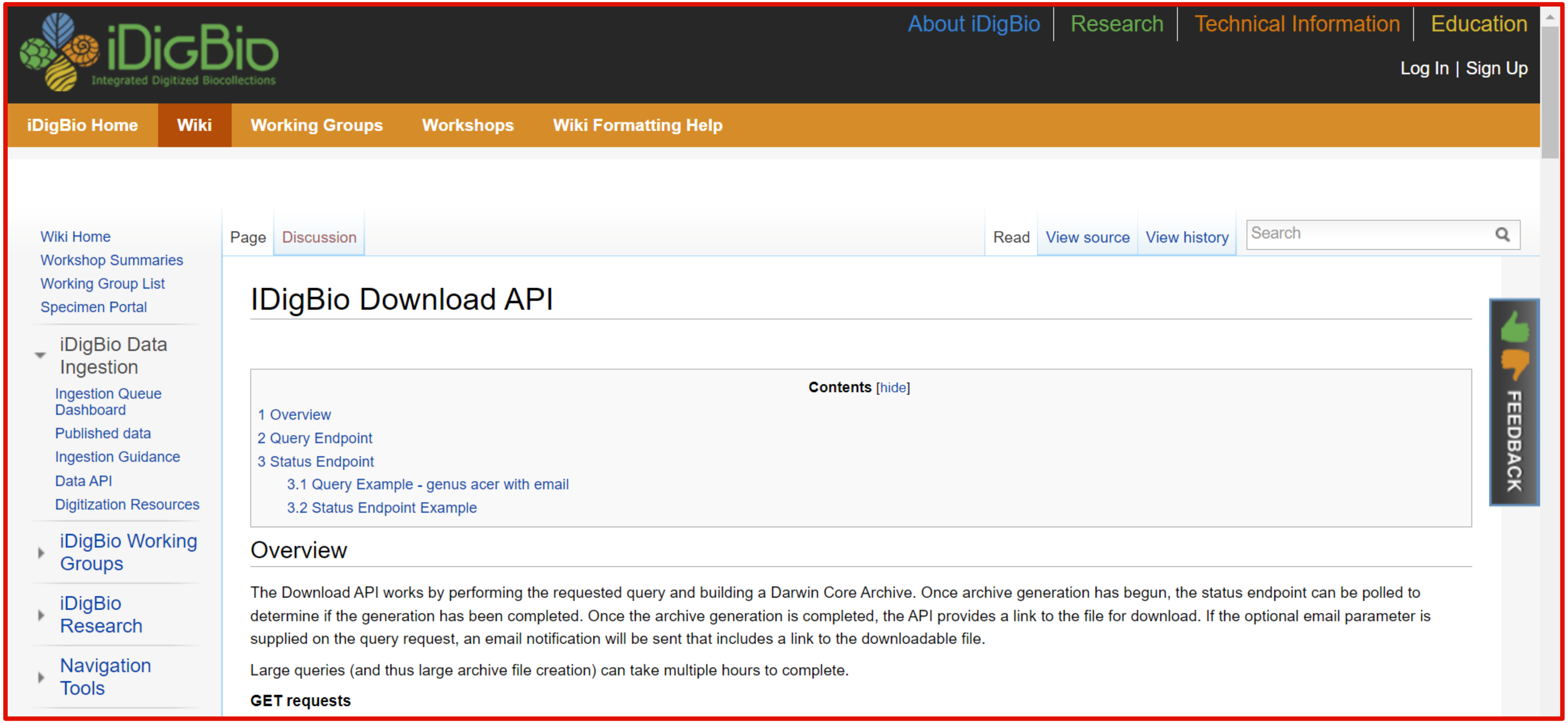
Download occurrence data from iDigBio

To download from iDigBio, full instructions are at the Download API reference.

To pull data from the command prompt, use the `curl` command to pull text response directly to terminal with the example query_url: Euphorbia

```
$ curl https://api.idigbio.org/v2/download/?rq=%7B%22genus%22%3A%22euphorbia%22%7D&email=donotreply%40idigbio.org
[1] 58979
astewart@murderbot:~/git/tutorials$ {
  "complete": false,
  "created": "2022-05-02T15:28:41.730968+00:00",
  "expires": "2022-06-01T15:28:41.628063+00:00",
  "hash": "18911492e8517926cb8693fc9f971cf066107016",
  "query": {
    "core_source": "indexterms",
    "core_type": "records",
    "form": "dwca-csv",
    "mediarecord_fields": null,
    "mq": null,
    "record_fields": null,
    "rq": {
      "genus": "euphorbia"
    }
  },
  "status_url": "https://api.idigbio.org/v2/download/d54c0ad7-6697-4096-9f11-b2a9a6041a38",
  "task_status": "PENDING"
}
```

Data Preparation: occurrence records



IDigBio Download API

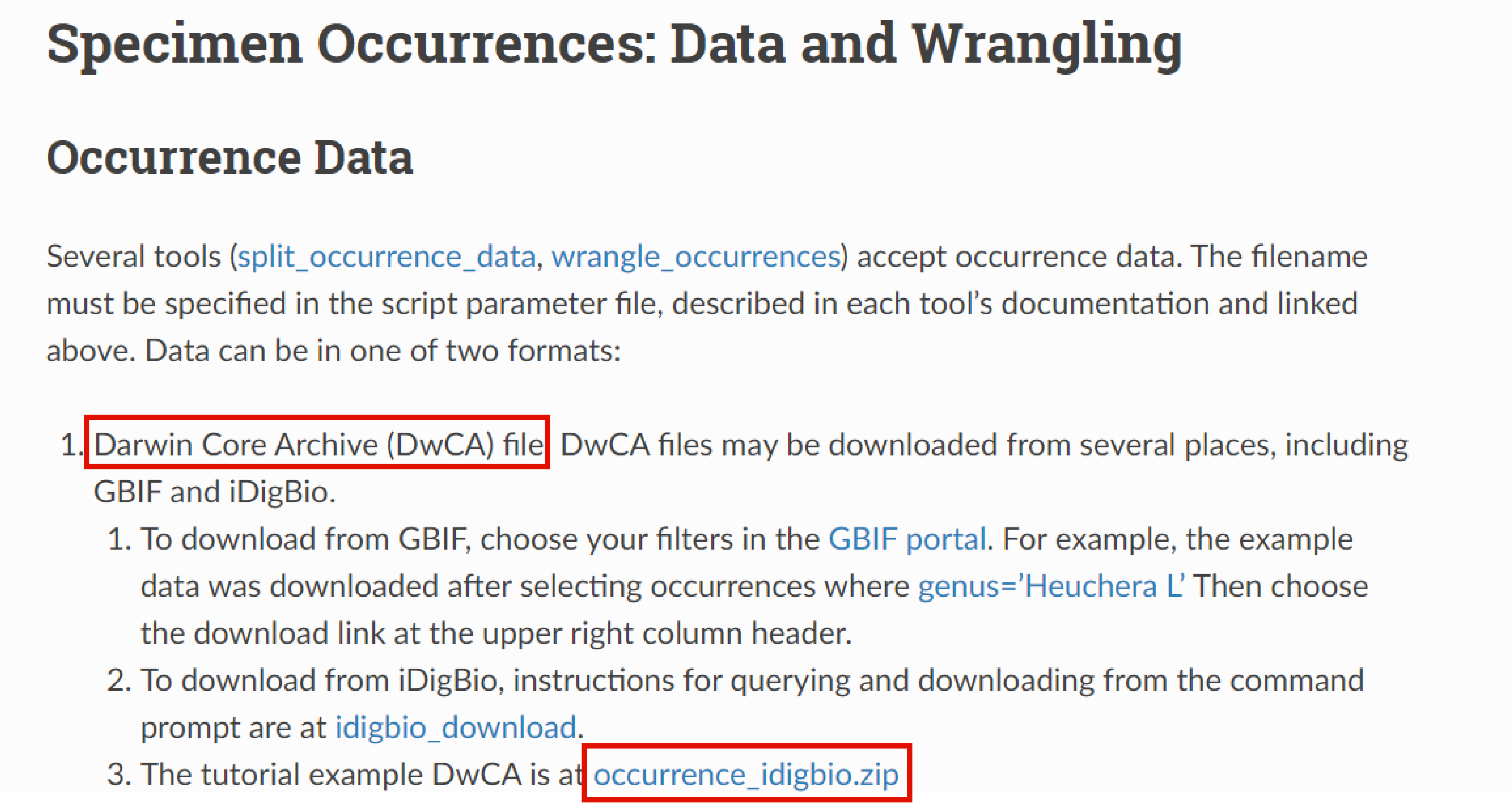
Overview

The Download API works by performing the requested query and building a Darwin Core Archive. Once archive generation has begun, the status endpoint can be polled to determine if the generation has been completed. Once the archive generation is completed, the API provides a link to the file for download. If the optional email parameter is supplied on the query request, an email notification will be sent that includes a link to the downloadable file.

Large queries (and thus large archive file creation) can take multiple hours to complete.

GET requests

Data Preparation: occurrence records



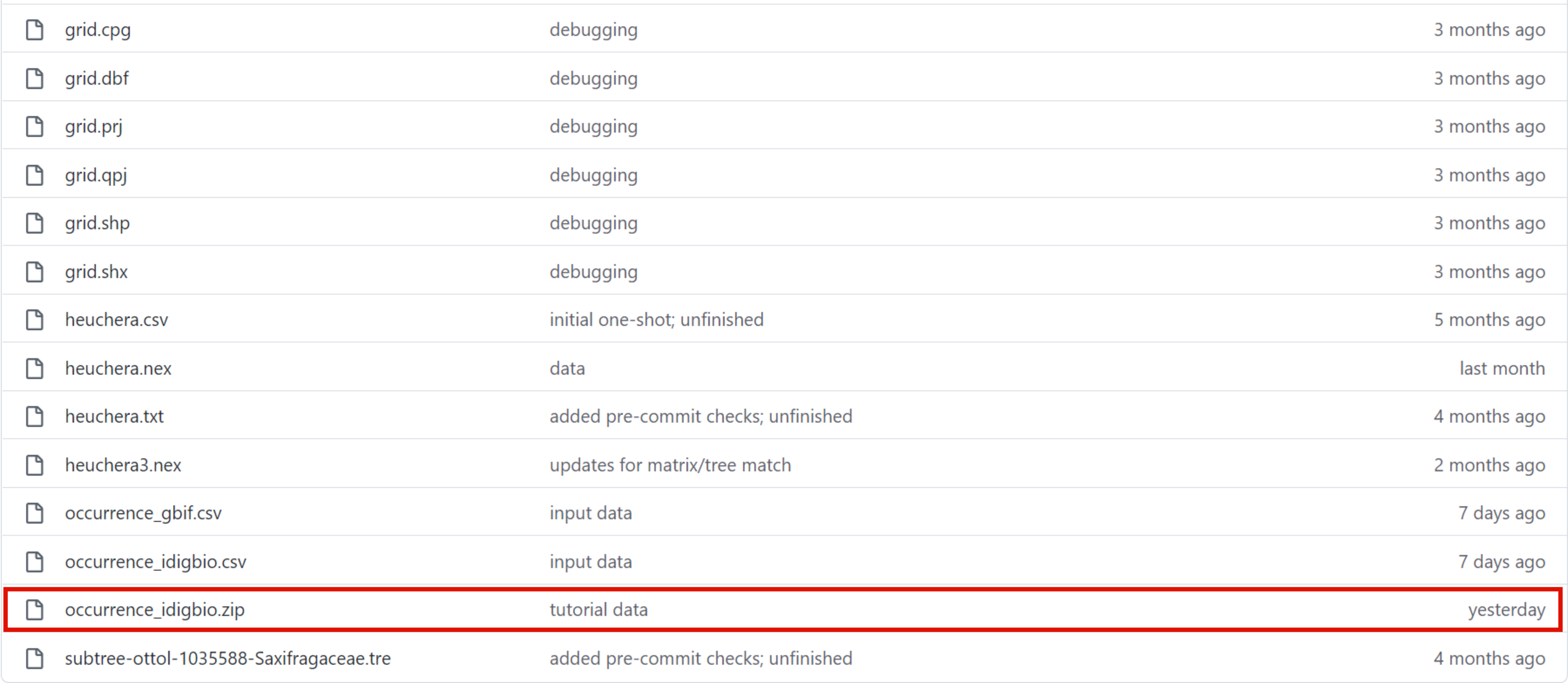
Specimen Occurrences: Data and Wrangling

Occurrence Data

Several tools (split_occurrence_data, wrangle_occurrences) accept occurrence data. The filename must be specified in the script parameter file, described in each tool’s documentation and linked above. Data can be in one of two formats:

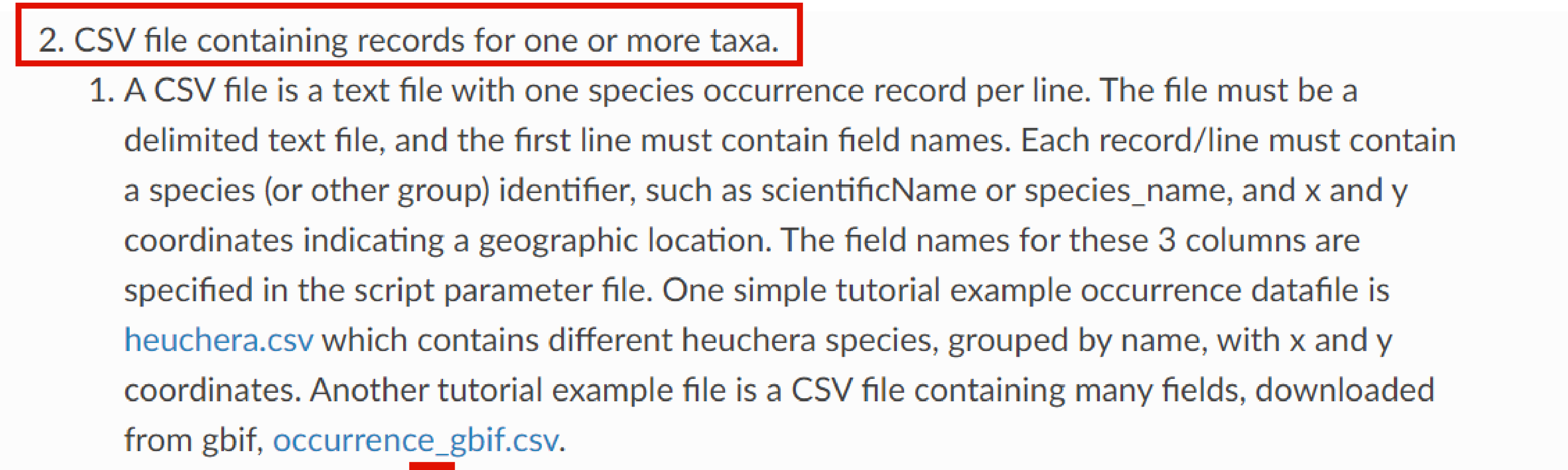
1. Darwin Core Archive (DwCA) file DwCA files may be downloaded from several places, including GBIF and iDigBio.
   1. To download from GBIF, choose your filters in the GBIF portal. For example, the example data was downloaded after selecting occurrences where genus=’Heuchera L’ Then choose the download link at the upper right column header.
   2. To download from iDigBio, instructions for querying and downloading from the command prompt are at idigbio_download.
   3. The tutorial example DwCA is at occurrence_idigbio.zip

Data Preparation: occurrence records

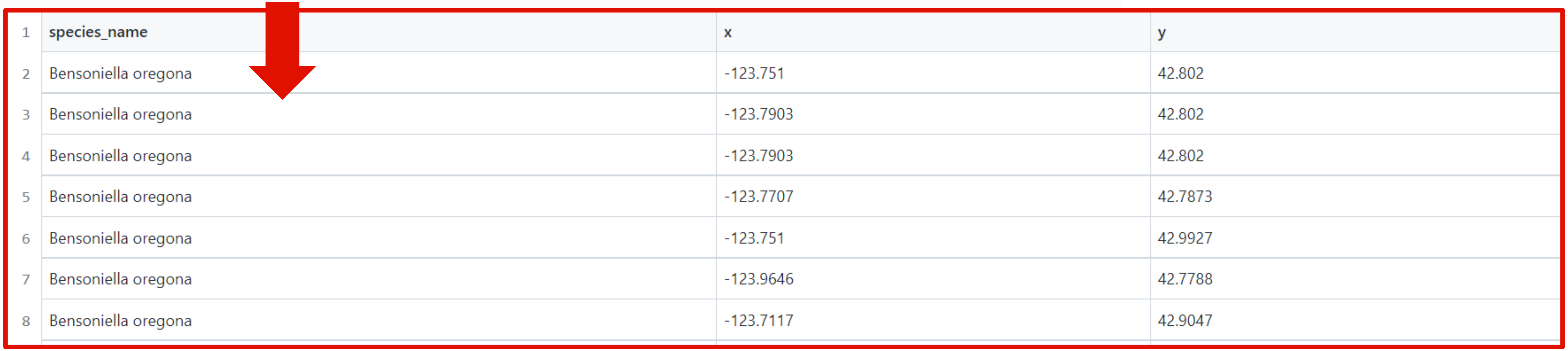


grid.cpg	debugging	3 months ago
grid.dbf	debugging	3 months ago
grid.prj	debugging	3 months ago
grid.qpj	debugging	3 months ago
grid.shp	debugging	3 months ago
grid.shx	debugging	3 months ago
heuchera.csv	initial one-shot; unfinished	5 months ago
heuchera.nex	data	last month
heuchera.txt	added pre-commit checks; unfinished	4 months ago
heuchera3.nex	updates for matrix/tree match	2 months ago
occurrence_gbif.csv	input data	7 days ago
occurrence_idigbio.csv	input data	7 days ago
occurrence_idigbio.zip	tutorial data	yesterday
subtree-ottol-1035588-Saxifragaceae.tre	added pre-commit checks; unfinished	4 months ago

Data Preparation: occurrence records

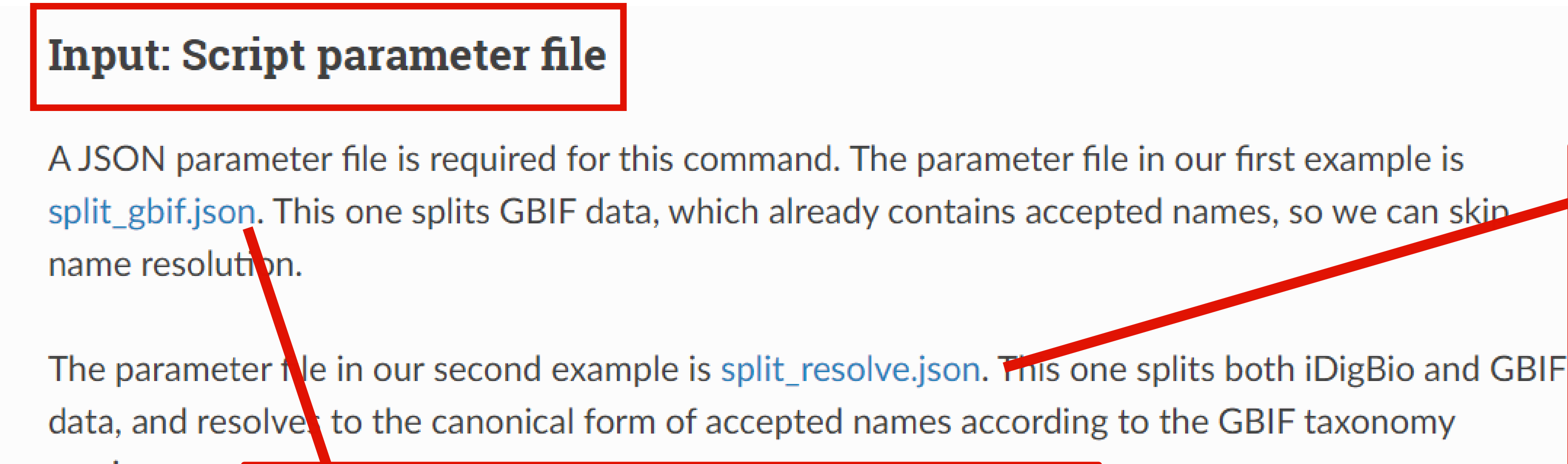


2. CSV file containing records for one or more taxa.
    1. A CSV file is a text file with one species occurrence record per line. The file must be a delimited text file, and the first line must contain field names. Each record/line must contain a species (or other group) identifier, such as scientificName or species_name, and x and y coordinates indicating a geographic location. The field names for these 3 columns are specified in the script parameter file. One simple tutorial example occurrence datafile is heuchera.csv which contains different heuchera species, grouped by name, with x and y coordinates. Another tutorial example file is a CSV file containing many fields, downloaded from gbif, occurrence_gbif.csv.



species_name	x	y
Bensoniella oregona	-123.751	42.802
Bensoniella oregona	-123.7903	42.802
Bensoniella oregona	-123.7903	42.802
Bensoniella oregona	-123.7707	42.7873
Bensoniella oregona	-123.751	42.9927
Bensoniella oregona	-123.9646	42.7788
Bensoniella oregona	-123.7117	42.9047

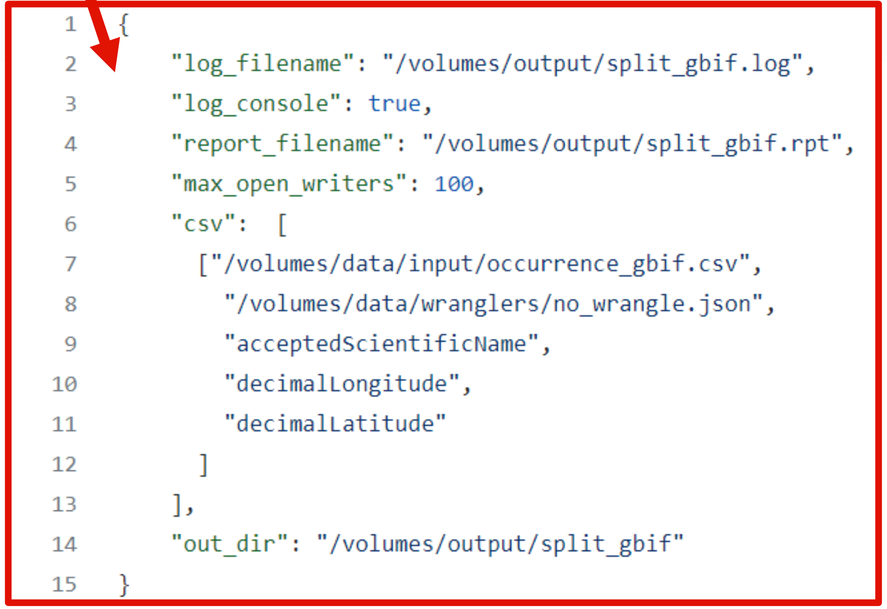
Data Preparation: script file



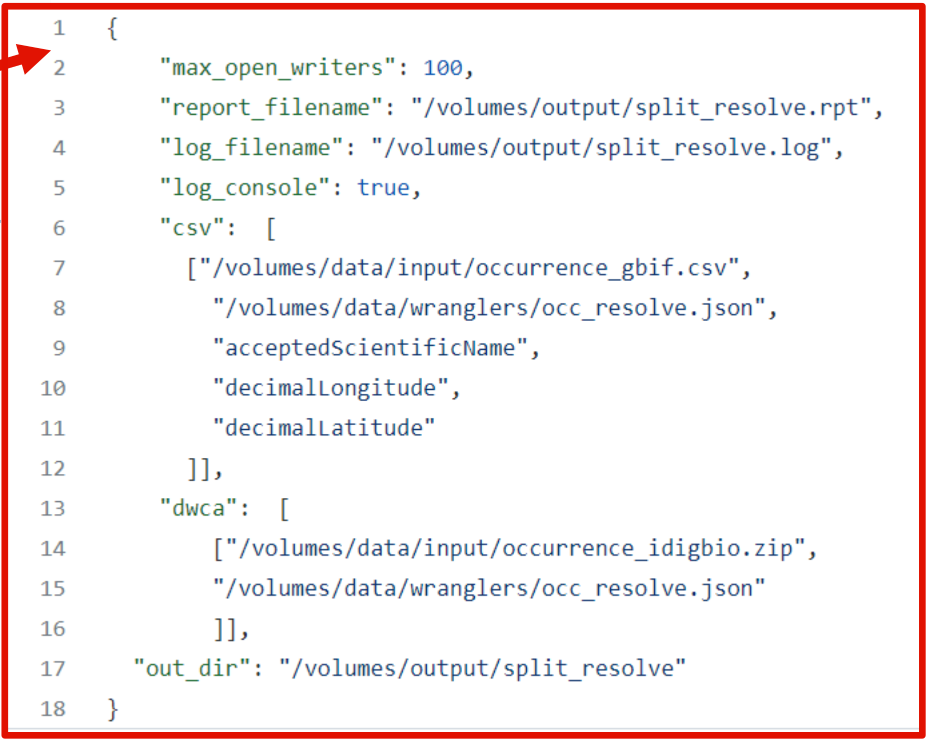
Input: Script parameter file

A JSON parameter file is required for this command. The parameter file in our first example is split_gbif.json. This one splits GBIF data, which already contains accepted names, so we can skip name resolution.

The parameter file in our second example is split_resolve.json. This one splits both iDigBio and GBIF data, and resolves to the canonical form of accepted names according to the GBIF taxonomy service.

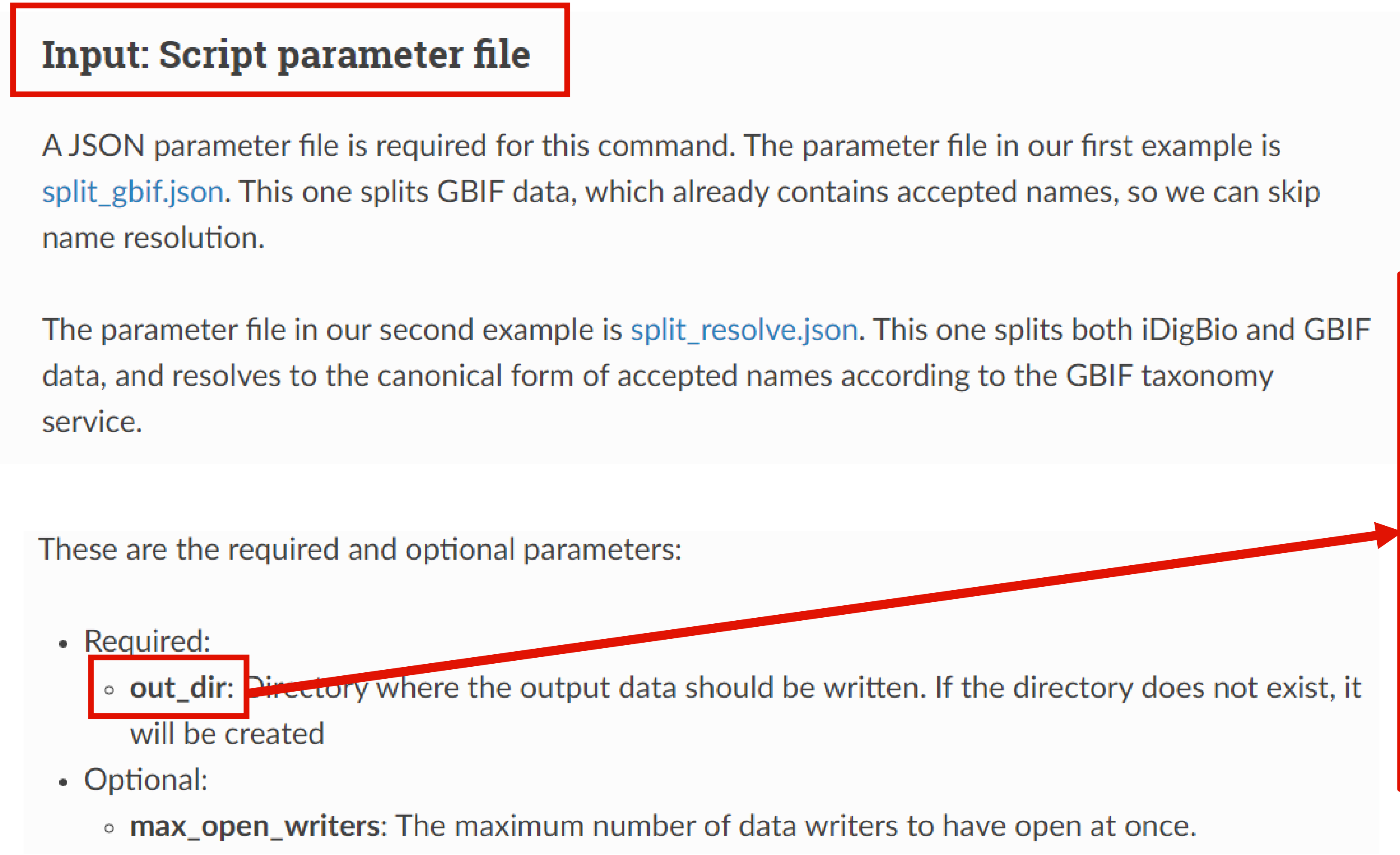


```
{
    "log_filename": "/volumes/output/split_gbif.log",
    "log_console": true,
    "report_filename": "/volumes/output/split_gbif.rpt",
    "max_open_writers": 100,
    "csv":  [
      ["/volumes/data/input/occurrence_gbif.csv",
        "/volumes/data/wranglers/no_wrangle.json",
        "acceptedScientificName",
        "decimalLongitude",
        "decimalLatitude"
      ]
    ],
    "out_dir": "/volumes/output/split_gbif"
}
```



```
{
    "max_open_writers": 100,
    "report_filename": "/volumes/output/split_resolve.rpt",
    "log_filename": "/volumes/output/split_resolve.log",
    "log_console": true,
    "csv":  [
      ["/volumes/data/input/occurrence_gbif.csv",
        "/volumes/data/wranglers/occ_resolve.json",
        "acceptedScientificName",
        "decimalLongitude",
        "decimalLatitude"
      ]],
    "dwca":  [
        ["/volumes/data/input/occurrence_idigbio.zip",
        "/volumes/data/wranglers/occ_resolve.json"
        ]],
  "out_dir": "/volumes/output/split_resolve"
}
```

Data Preparation: script file



Input: Script parameter file

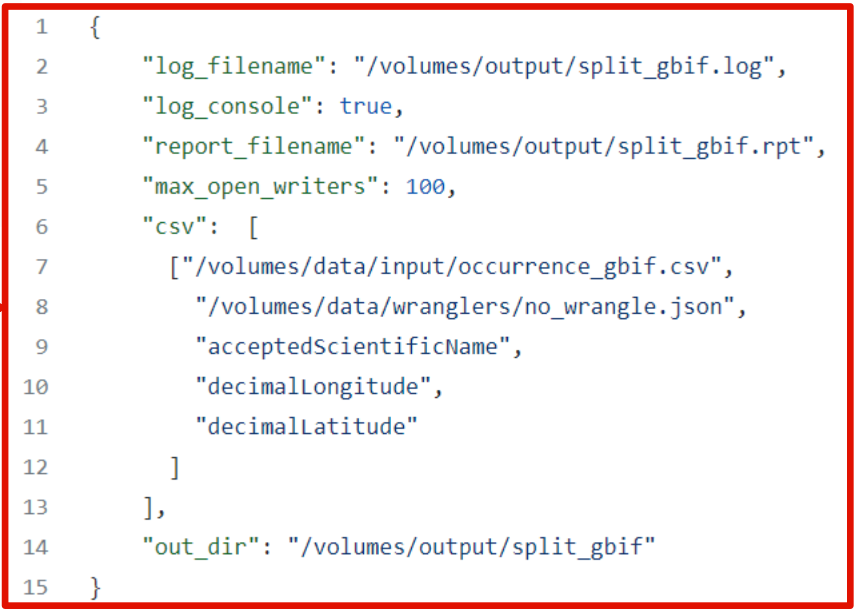
A JSON parameter file is required for this command. The parameter file in our first example is split_gbif.json. This one splits GBIF data, which already contains accepted names, so we can skip name resolution.

The parameter file in our second example is split_resolve.json. This one splits both iDigBio and GBIF data, and resolves to the canonical form of accepted names according to the GBIF taxonomy service.

These are the required and optional parameters:

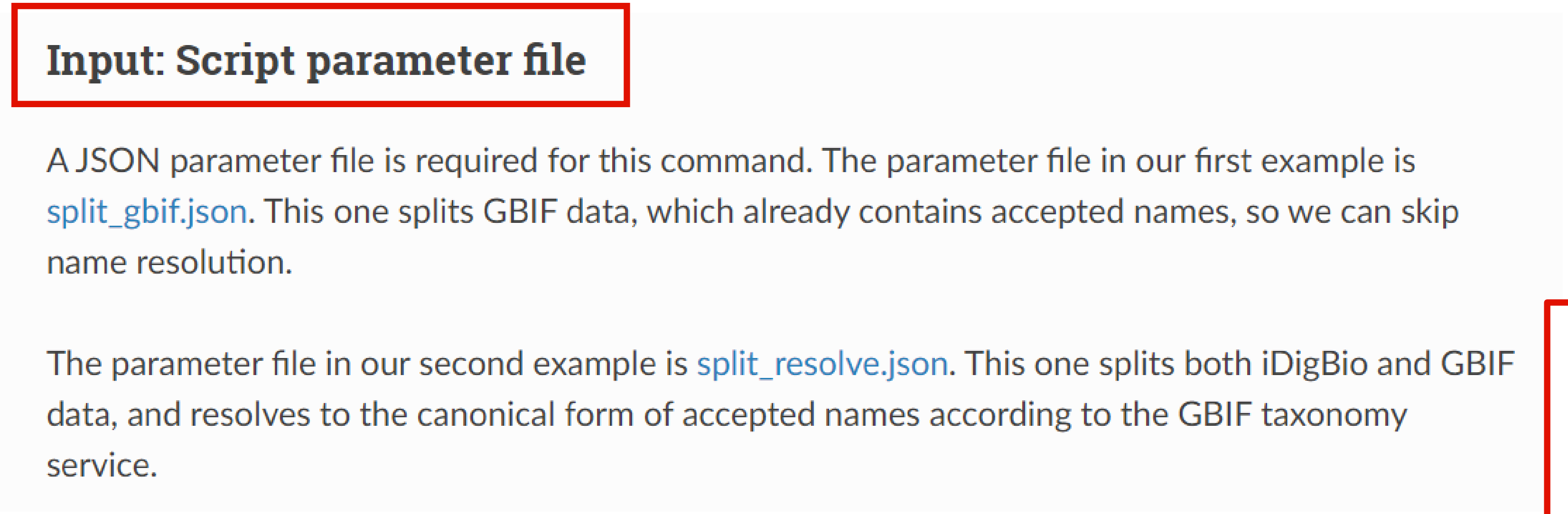
- Required:
  - **out_dir**: Directory where the output data should be written. If the directory does not exist, it will be created
- Optional:
  - **max_open_writers**: The maximum number of data writers to have open at once.

Required parameter



```
{
    "log_filename": "/volumes/output/split_gbif.log",
    "log_console": true,
    "report_filename": "/volumes/output/split_gbif.rpt",
    "max_open_writers": 100,
    "csv":  [
      ["/volumes/data/input/occurrence_gbif.csv",
        "/volumes/data/wranglers/no_wrangle.json",
        "acceptedScientificName",
        "decimalLongitude",
        "decimalLatitude"
      ]
    ],
    "out_dir": "/volumes/output/split_gbif"
}
```

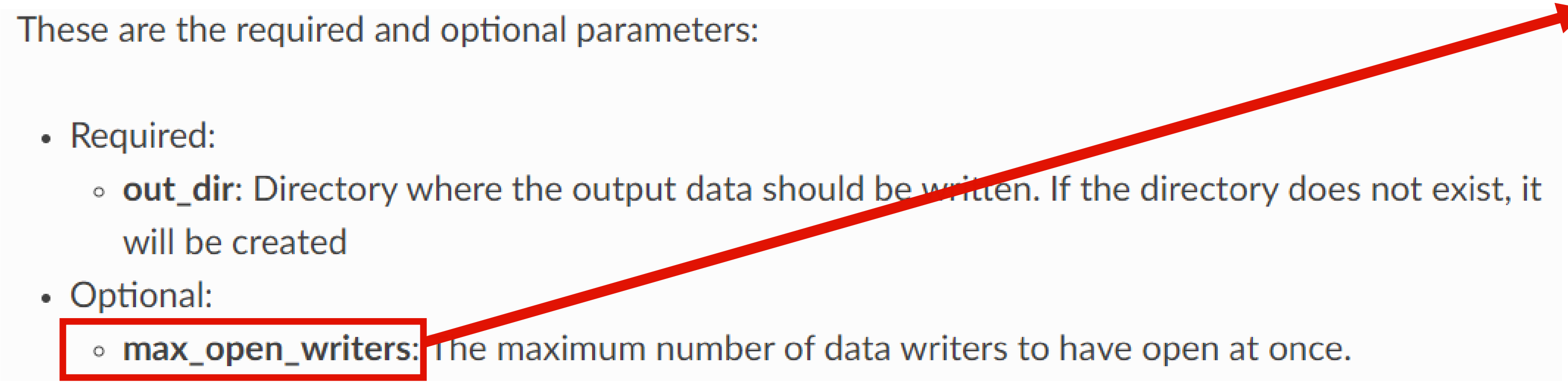
Data Preparation: script file



Input: Script parameter file

A JSON parameter file is required for this command. The parameter file in our first example is split_gbif.json. This one splits GBIF data, which already contains accepted names, so we can skip name resolution.

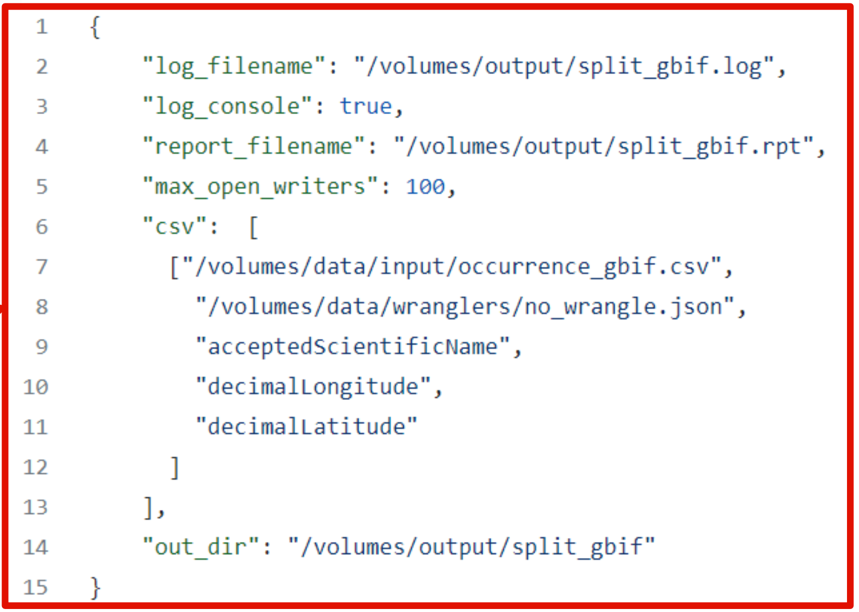
The parameter file in our second example is split_resolve.json. This one splits both iDigBio and GBIF data, and resolves to the canonical form of accepted names according to the GBIF taxonomy service.



These are the required and optional parameters:

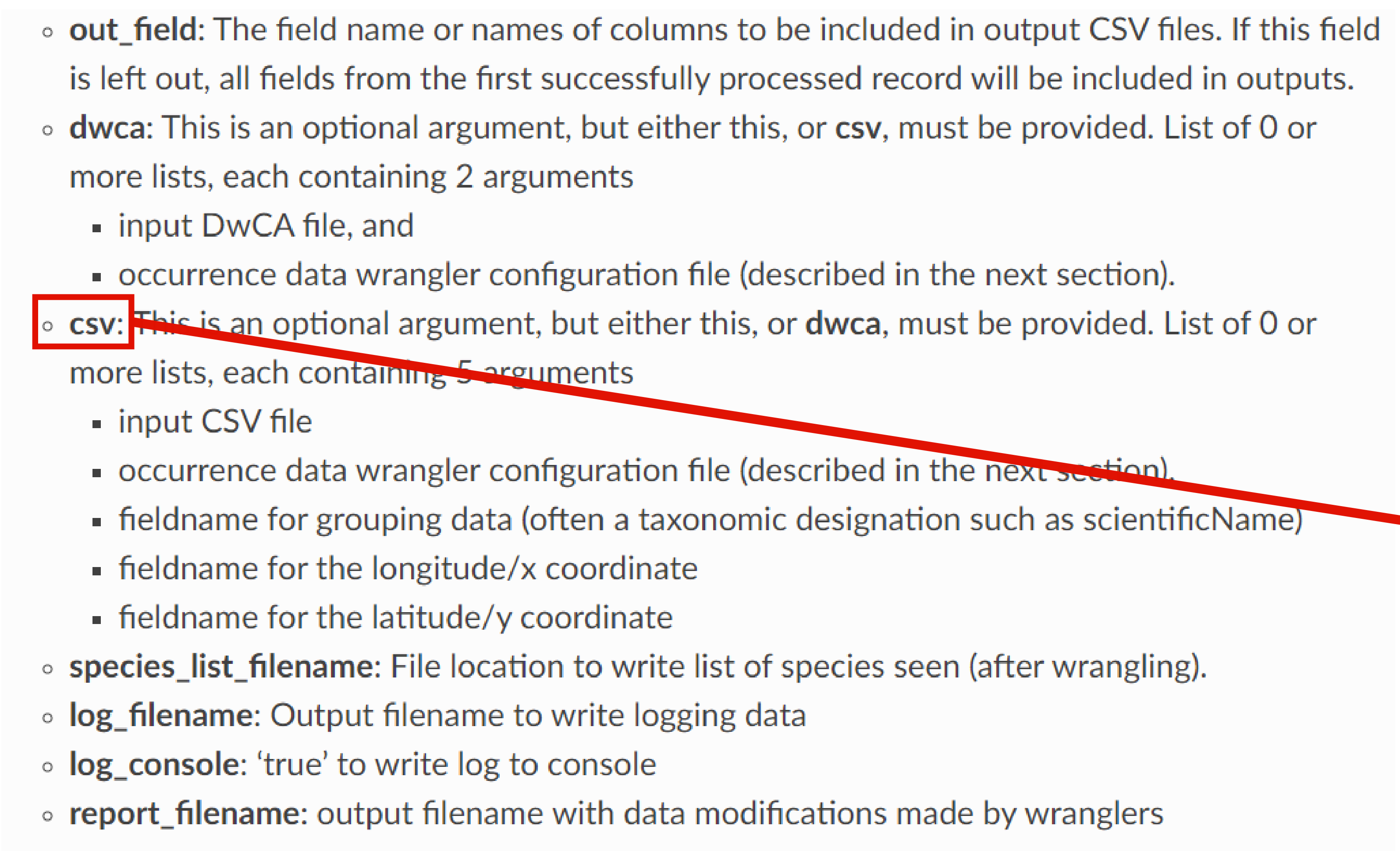
- Required:
  - **out_dir**: Directory where the output data should be written. If the directory does not exist, it will be created
- Optional:
  - **max_open_writers**: The maximum number of data writers to have open at once.

Optional parameters used



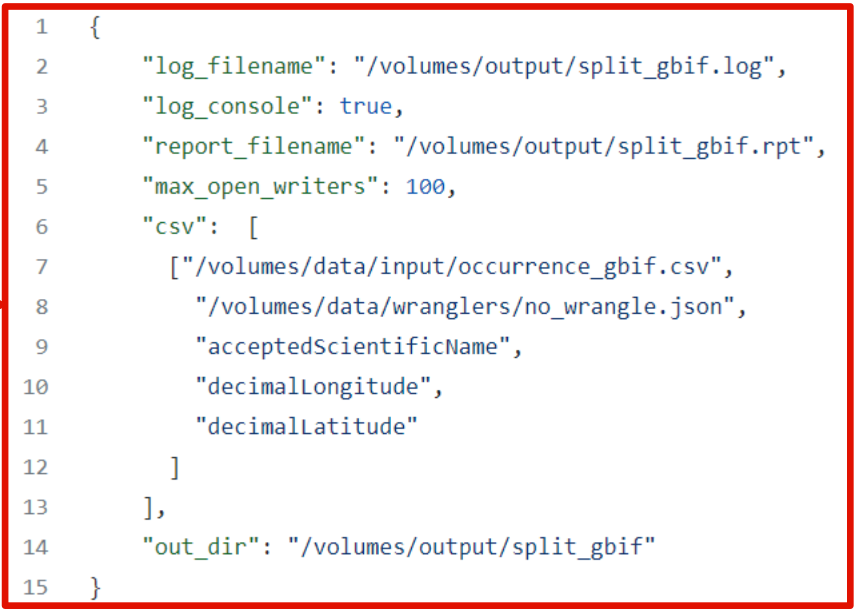
```
{
    "log_filename": "/volumes/output/split_gbif.log",
    "log_console": true,
    "report_filename": "/volumes/output/split_gbif.rpt",
    "max_open_writers": 100,
    "csv":  [
      ["/volumes/data/input/occurrence_gbif.csv",
        "/volumes/data/wranglers/no_wrangle.json",
        "acceptedScientificName",
        "decimalLongitude",
        "decimalLatitude"
      ]
    ],
    "out_dir": "/volumes/output/split_gbif"
}
```

Data Preparation: script file



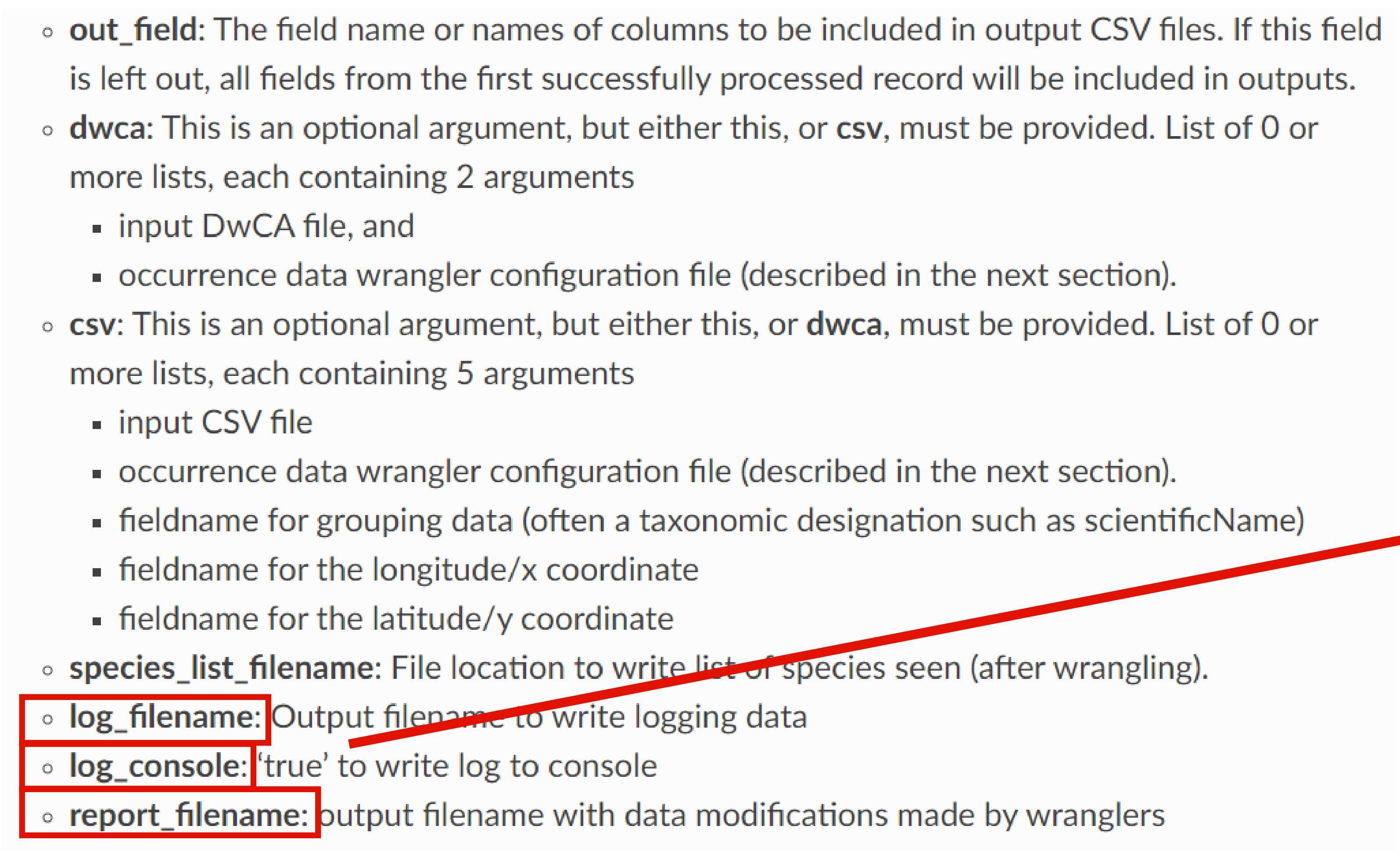
- **out_field:** The field name or names of columns to be included in output CSV files. If this field is left out, all fields from the first successfully processed record will be included in outputs.
- **dwca**: This is an optional argument, but either this, or **csv**, must be provided. List of 0 or more lists, each containing 2 arguments
  - input DwCA file, and
  - occurrence data wrangler configuration file (described in the next section).
- **csv**: This is an optional argument, but either this, or **dwca**, must be provided. List of 0 or more lists, each containing 5 arguments
  - input CSV file
  - occurrence data wrangler configuration file (described in the next section).
  - fieldname for grouping data (often a taxonomic designation such as scientificName)
  - fieldname for the longitude/x coordinate
  - fieldname for the latitude/y coordinate
- **species_list_filename:** File location to write list of species seen (after wrangling).
- **log_filename**: Output filename to write logging data
- **log_console**: 'true' to write log to console
- **report_filename:** output filename with data modifications made by wranglers

Optional parameters used



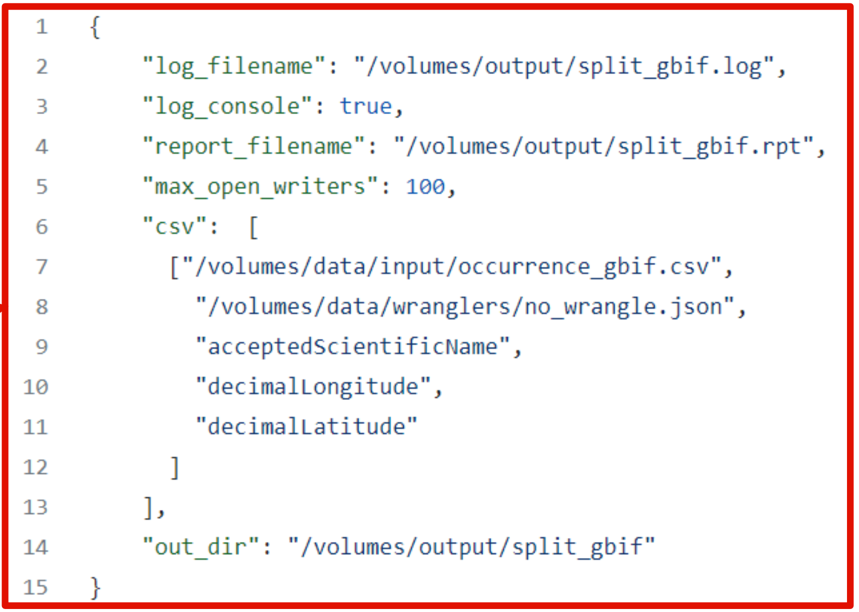
```
{
    "log_filename": "/volumes/output/split_gbif.log",
    "log_console": true,
    "report_filename": "/volumes/output/split_gbif.rpt",
    "max_open_writers": 100,
    "csv":  [
      ["/volumes/data/input/occurrence_gbif.csv",
        "/volumes/data/wranglers/no_wrangle.json",
        "acceptedScientificName",
        "decimalLongitude",
        "decimalLatitude"
      ]
    ],
    "out_dir": "/volumes/output/split_gbif"
}
```

Data Preparation: script file



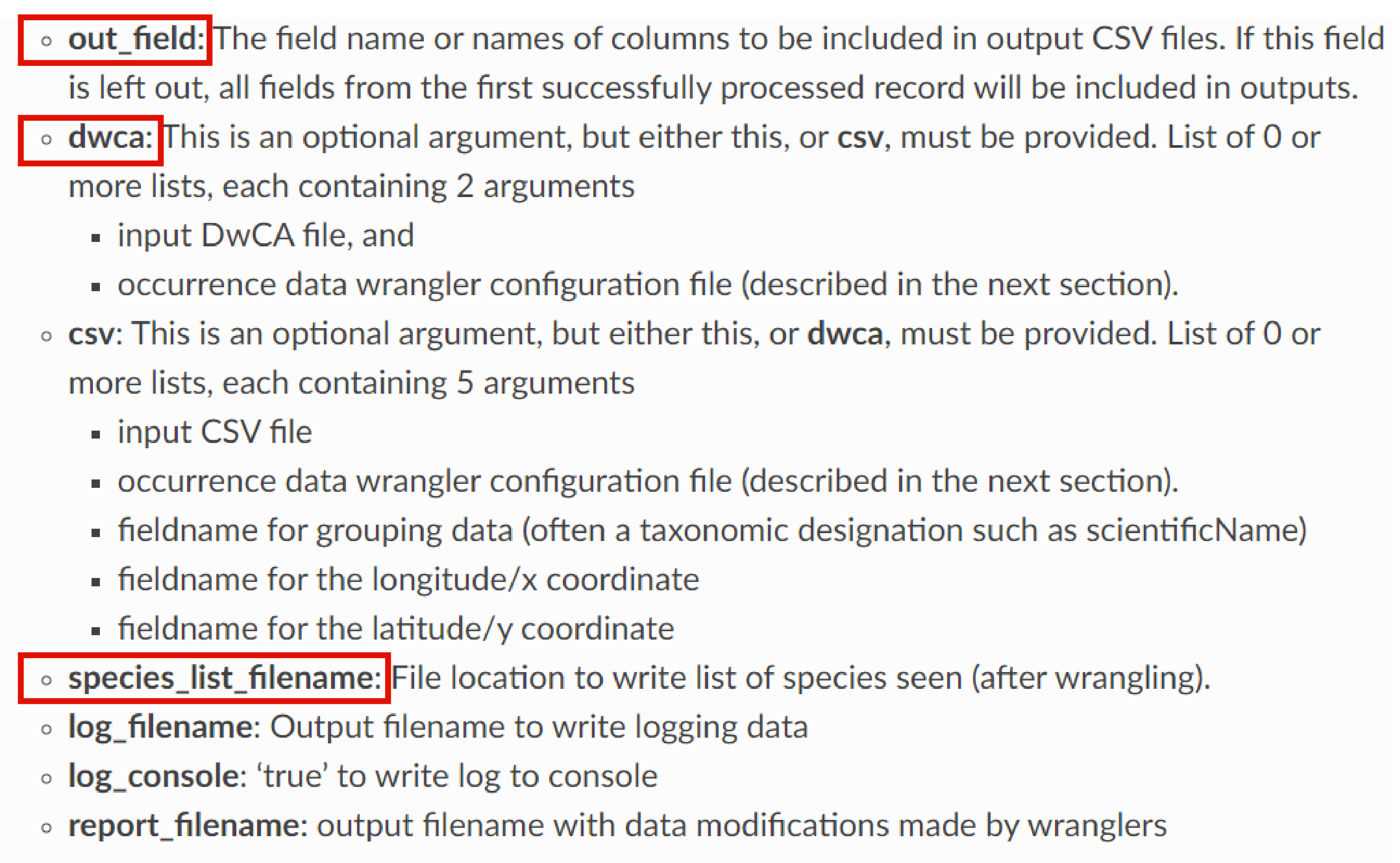
- **out_field:** The field name or names of columns to be included in output CSV files. If this field is left out, all fields from the first successfully processed record will be included in outputs.
- **dwca**: This is an optional argument, but either this, or **csv**, must be provided. List of 0 or more lists, each containing 2 arguments
  - input DwCA file, and
  - occurrence data wrangler configuration file (described in the next section).
- **csv**: This is an optional argument, but either this, or **dwca**, must be provided. List of 0 or more lists, each containing 5 arguments
  - input CSV file
  - occurrence data wrangler configuration file (described in the next section).
  - fieldname for grouping data (often a taxonomic designation such as scientificName)
  - fieldname for the longitude/x coordinate
  - fieldname for the latitude/y coordinate
- **species_list_filename:** File location to write list of species seen (after wrangling).
- **log_filename:** Output filename to write logging data
- **log_console:** 'true' to write log to console
- **report_filename:** output filename with data modifications made by wranglers

Optional parameters used



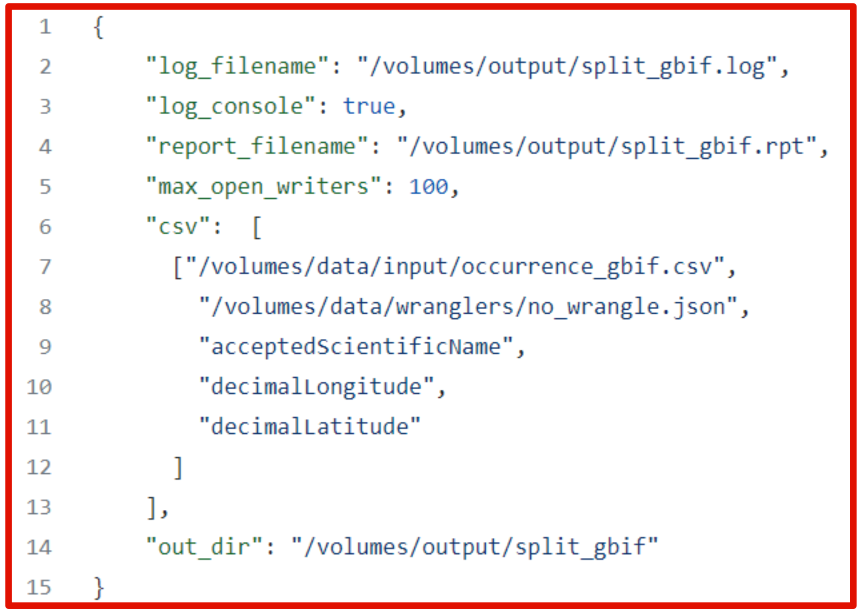
```
{
    "log_filename": "/volumes/output/split_gbif.log",
    "log_console": true,
    "report_filename": "/volumes/output/split_gbif.rpt",
    "max_open_writers": 100,
    "csv":  [
      ["/volumes/data/input/occurrence_gbif.csv",
        "/volumes/data/wranglers/no_wrangle.json",
        "acceptedScientificName",
        "decimalLongitude",
        "decimalLatitude"
      ]
    ],
    "out_dir": "/volumes/output/split_gbif"
}
```

Data Preparation: script file



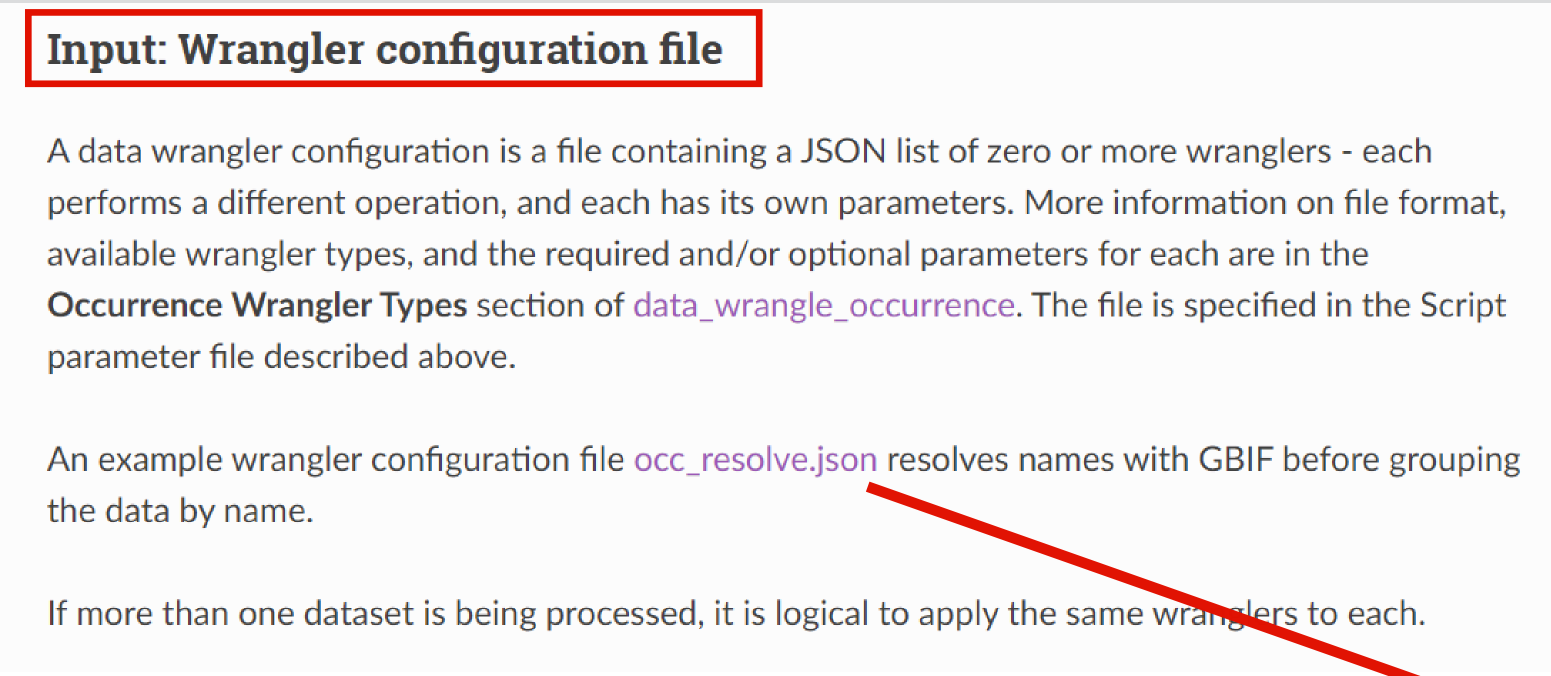
- **out_field:** The field name or names of columns to be included in output CSV files. If this field is left out, all fields from the first successfully processed record will be included in outputs.
- **dwca:** This is an optional argument, but either this, or **csv**, must be provided. List of 0 or more lists, each containing 2 arguments
  - input DwCA file, and
  - occurrence data wrangler configuration file (described in the next section).
- **csv**: This is an optional argument, but either this, or **dwca**, must be provided. List of 0 or more lists, each containing 5 arguments
  - input CSV file
  - occurrence data wrangler configuration file (described in the next section).
  - fieldname for grouping data (often a taxonomic designation such as scientificName)
  - fieldname for the longitude/x coordinate
  - fieldname for the latitude/y coordinate
- **species_list_filename:** File location to write list of species seen (after wrangling).
- **log_filename**: Output filename to write logging data
- **log_console**: ‘true’ to write log to console
- **report_filename**: output filename with data modifications made by wranglers

Optional parameters NOT used



```
{
    "log_filename": "/volumes/output/split_gbif.log",
    "log_console": true,
    "report_filename": "/volumes/output/split_gbif.rpt",
    "max_open_writers": 100,
    "csv":  [
      ["/volumes/data/input/occurrence_gbif.csv",
        "/volumes/data/wranglers/no_wrangle.json",
        "acceptedScientificName",
        "decimalLongitude",
        "decimalLatitude"
      ]
    ],
    "out_dir": "/volumes/output/split_gbif"
}
```

Data Preparation: wrangler file

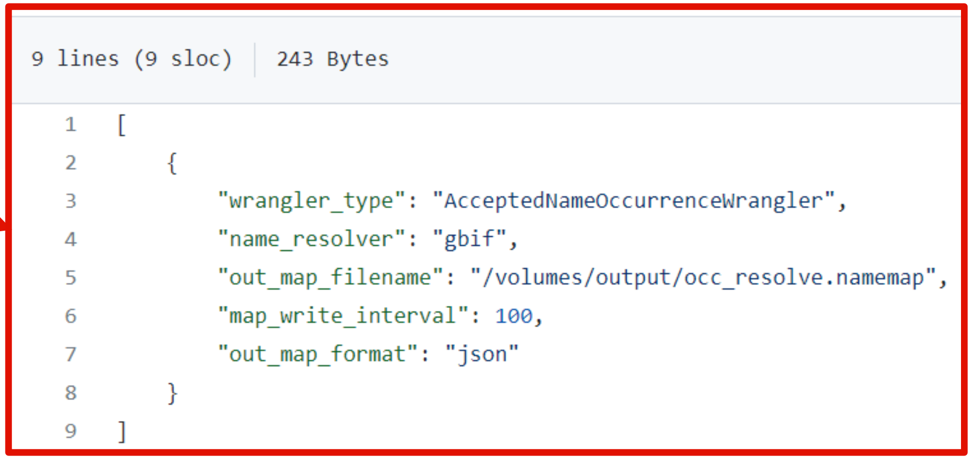


Input: Wrangler configuration file

A data wrangler configuration is a file containing a JSON list of zero or more wranglers - each performs a different operation, and each has its own parameters. More information on file format, available wrangler types, and the required and/or optional parameters for each are in the **Occurrence Wrangler Types** section of data_wrangle_occurrence. The file is specified in the Script parameter file described above.

An example wrangler configuration file occ_resolve.json resolves names with GBIF before grouping the data by name.

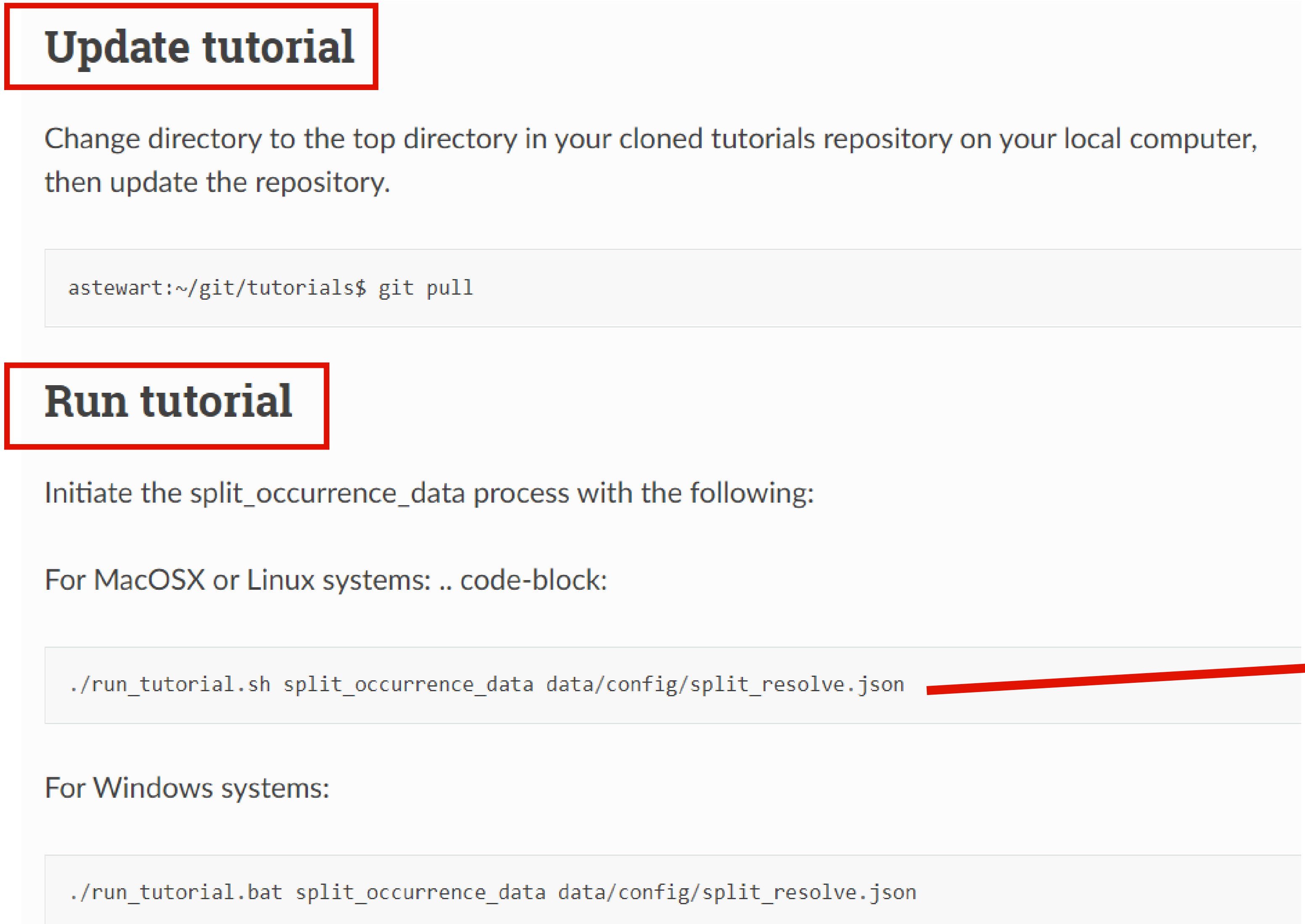
If more than one dataset is being processed, it is logical to apply the same wranglers to each.



9 lines (9 sloc) | 243 Bytes

```
[
    {
        "wrangler_type": "AcceptedNameOccurrenceWrangler",
        "name_resolver": "gbif",
        "out_map_filename": "/volumes/output/occ_resolve.namemap",
        "map_write_interval": 100,
        "out_map_format": "json"
    }
]
```

Let's run this tutorial!



Update tutorial

Change directory to the top directory in your cloned tutorials repository on your local computer, then update the repository.

```
astewart:~/git/tutorials$ git pull
```

Run tutorial

Initiate the split_occurrence_data process with the following:

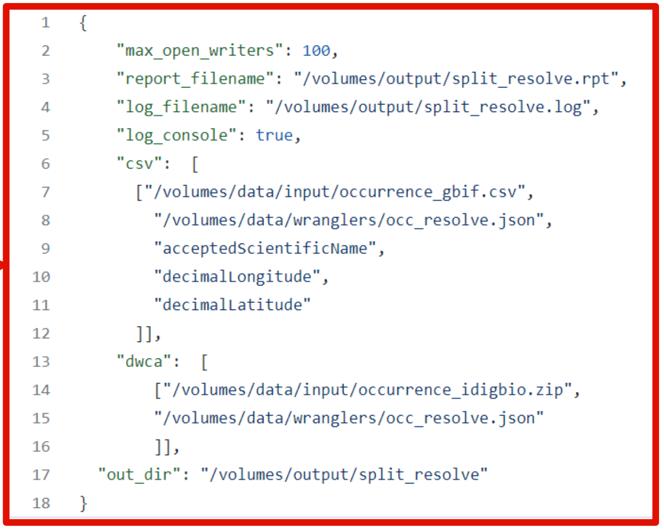
For MacOSX or Linux systems: .. code-block:

```
./run_tutorial.sh split_occurrence_data data/config/split_resolve.json
```

For Windows systems:

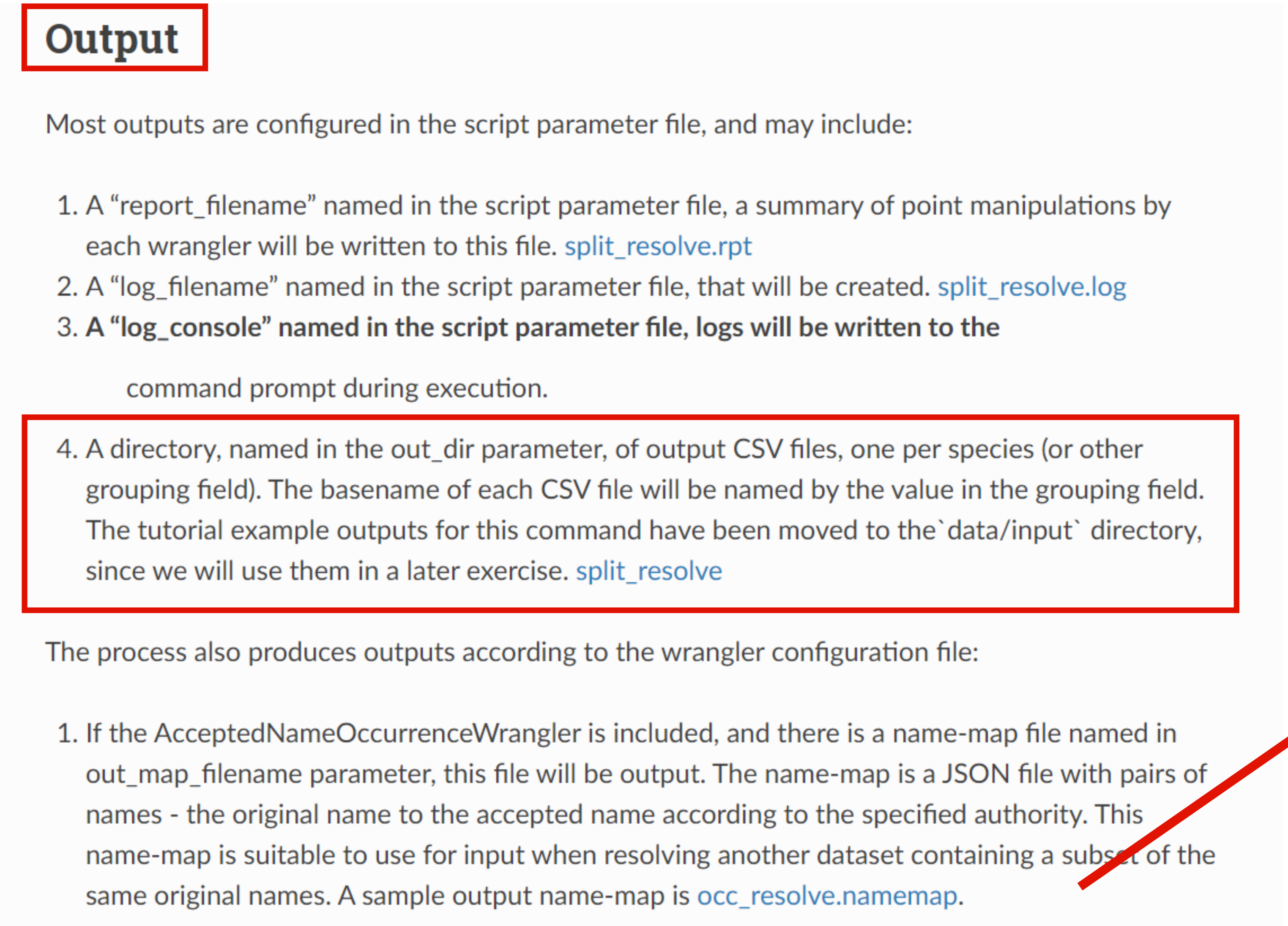
```
./run_tutorial.bat split_occurrence_data data/config/split_resolve.json
```

Goal: (usually) produce a file per species containing occurrence data. SPLITTING FILES IS ESSENTIAL FOR FACILITATING SDMS!



```
{
  "max_open_writers": 100,
  "report_filename": "/volumes/output/split_resolve.rpt",
  "log_filename": "/volumes/output/split_resolve.log",
  "log_console": true,
  "csv":  [
   ["/volumes/data/input/occurrence_gbif.csv",
    "/volumes/data/wranglers/occ_resolve.json",
    "acceptedScientificName",
    "decimalLongitude",
    "decimalLatitude"
   ]],
  "dwca":  [
    ["/volumes/data/input/occurrence_idigbio.zip",
    "/volumes/data/wranglers/occ_resolve.json"
    ]],
 "out_dir": "/volumes/output/split_resolve"
}
```

Let’s look at the output!



Output

Most outputs are configured in the script parameter file, and may include:

1. A “report_filename” named in the script parameter file, a summary of point manipulations by each wrangler will be written to this file. split_resolve.rpt
2. A “log_filename” named in the script parameter file, that will be created. split_resolve.log
3. **A “log_console” named in the script parameter file, logs will be written to the**

    command prompt during execution.

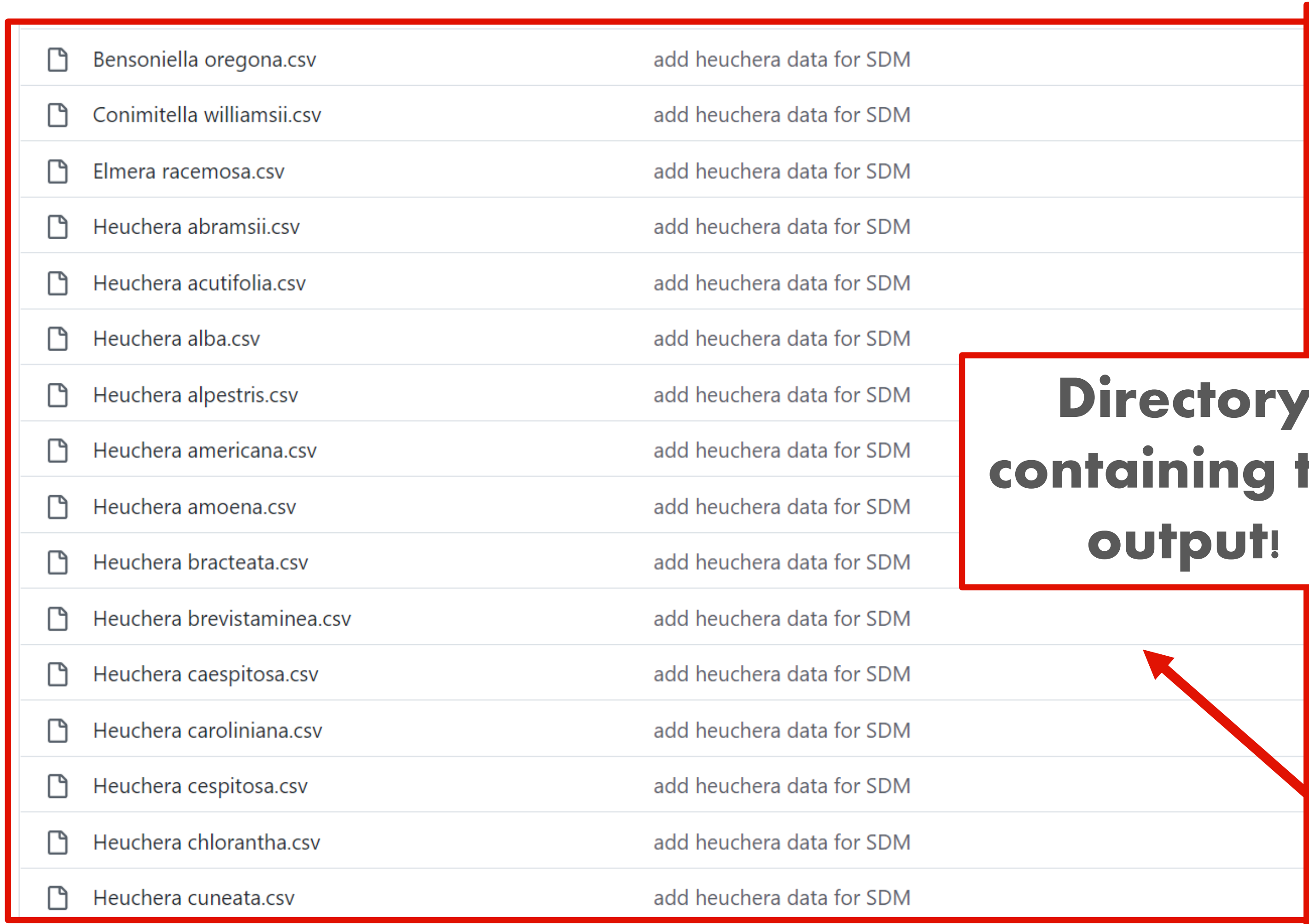
4. A directory, named in the out_dir parameter, of output CSV files, one per species (or other grouping field). The basename of each CSV file will be named by the value in the grouping field. The tutorial example outputs for this command have been moved to the`data/input` directory, since we will use them in a later exercise. split_resolve

The process also produces outputs according to the wrangler configuration file:

1. If the AcceptedNameOccurrenceWrangler is included, and there is a name-map file named in out_map_filename parameter, this file will be output. The name-map is a JSON file with pairs of names - the original name to the accepted name according to the specified authority. This name-map is suitable to use for input when resolving another dataset containing a subset of the same original names. A sample output name-map is occ_resolve.namemap.

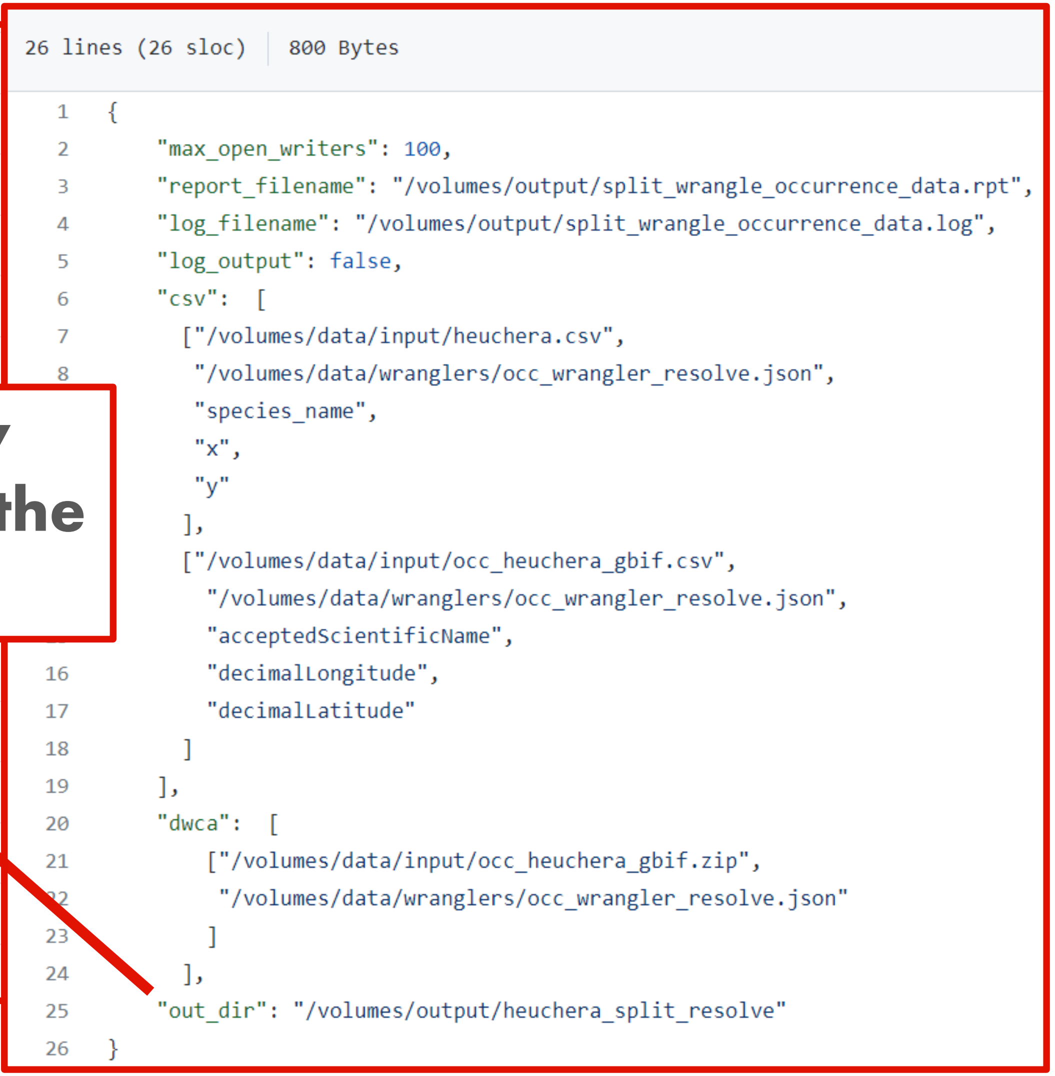
WE SAW THIS AS AN OUTPUT WHEN RESOLVING NAMES!

Let’s look at the output!



Bensoniella oregona.csv	add heuchera data for SDM
Conimitella williamsii.csv	add heuchera data for SDM
Elmera racemosa.csv	add heuchera data for SDM
Heuchera abramsii.csv	add heuchera data for SDM
Heuchera acutifolia.csv	add heuchera data for SDM
Heuchera alba.csv	add heuchera data for SDM
Heuchera alpestris.csv	add heuchera data for SDM
Heuchera americana.csv	add heuchera data for SDM
Heuchera amoena.csv	add heuchera data for SDM
Heuchera bracteata.csv	add heuchera data for SDM
Heuchera brevistaminea.csv	add heuchera data for SDM
Heuchera caespitosa.csv	add heuchera data for SDM
Heuchera caroliniana.csv	add heuchera data for SDM
Heuchera cespitosa.csv	add heuchera data for SDM
Heuchera chlorantha.csv	add heuchera data for SDM
Heuchera cuneata.csv	add heuchera data for SDM

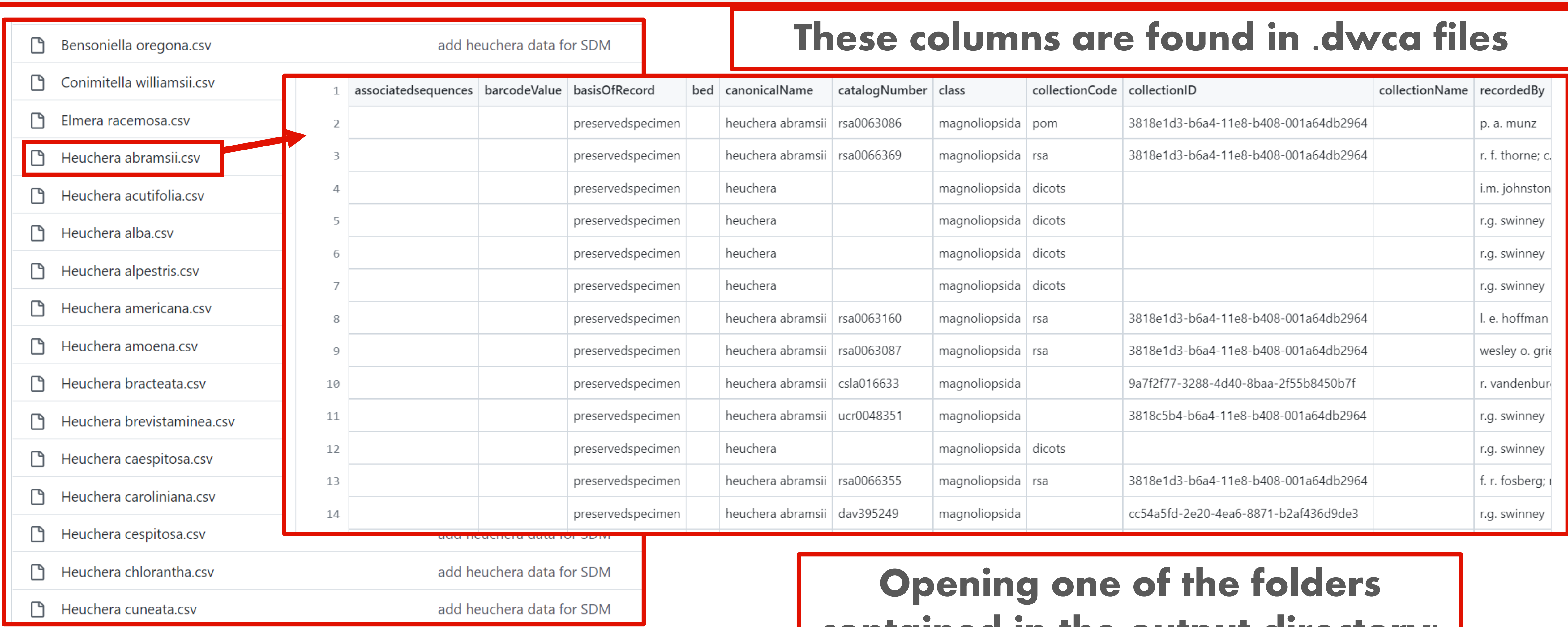
Directory containing the output!



26 lines (26 sloc) | 800 Bytes

```
{
    "max_open_writers": 100,
    "report_filename": "/volumes/output/split_wrangle_occurrence_data.rpt",
    "log_filename": "/volumes/output/split_wrangle_occurrence_data.log",
    "log_output": false,
    "csv":  [
      ["/volumes/data/input/heuchera.csv",
       "/volumes/data/wranglers/occ_wrangler_resolve.json",
       "species_name",
       "x",
       "y"
      ],
      ["/volumes/data/input/occ_heuchera_gbif.csv",
        "/volumes/data/wranglers/occ_wrangler_resolve.json",
        "acceptedScientificName",
        "decimalLongitude",
        "decimalLatitude"
      ]
    ],
    "dwca":  [
        ["/volumes/data/input/occ_heuchera_gbif.zip",
         "/volumes/data/wranglers/occ_wrangler_resolve.json"
        ]
      ],
    "out_dir": "/volumes/output/heuchera_split_resolve"
}
```

Let’s look at the output!



Bensoniella oregona.csv — add heuchera data for SDM
Conimitella williamsii.csv
Elmera racemosa.csv
Heuchera abramsii.csv
Heuchera acutifolia.csv
Heuchera alba.csv
Heuchera alpestris.csv
Heuchera americana.csv
Heuchera amoena.csv
Heuchera bracteata.csv
Heuchera brevistaminea.csv
Heuchera caespitosa.csv
Heuchera caroliniana.csv
Heuchera cespitosa.csv
Heuchera chlorantha.csv — add heuchera data for SDM
Heuchera cuneata.csv — add heuchera data for SDM

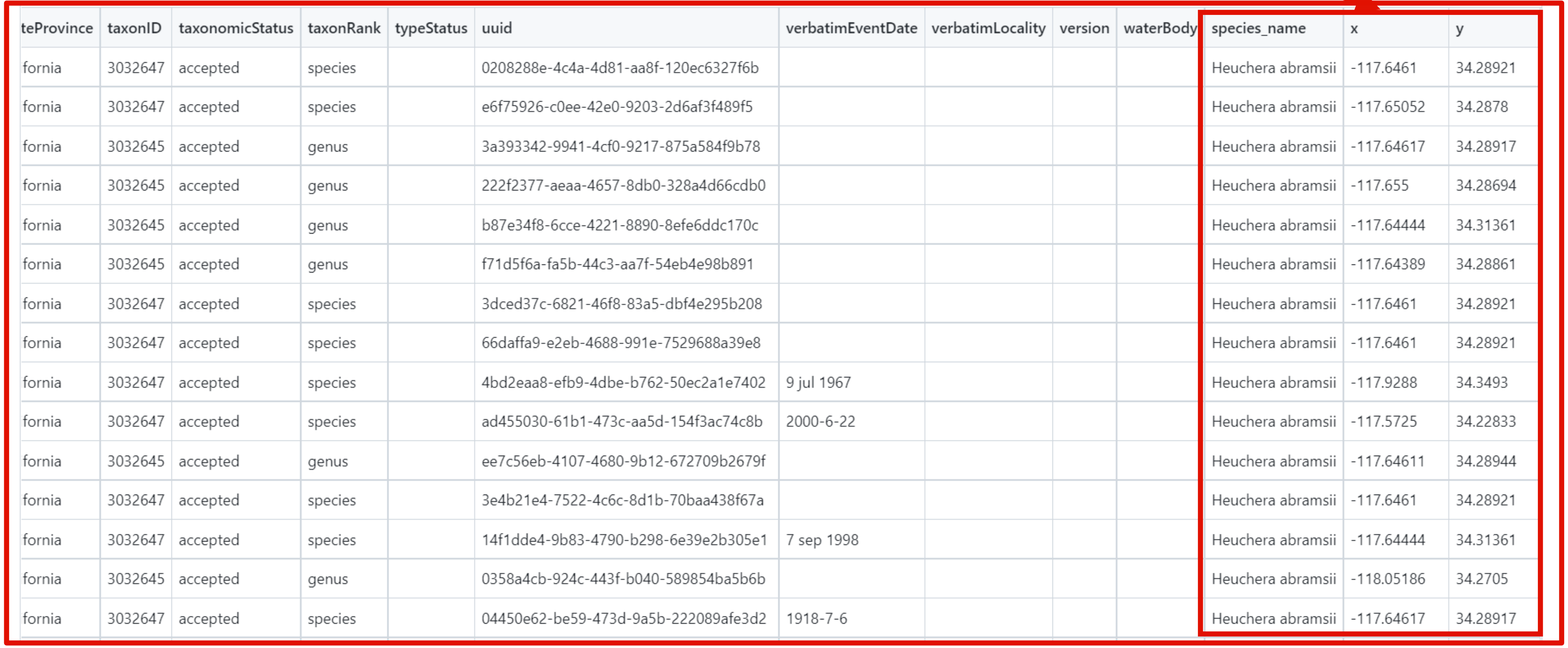
associatedsequences	barcodeValue	basisOfRecord	bed	canonicalName	catalogNumber	class	collectionCode	collectionID	collectionName	recordedBy
		preservedspecimen		heuchera abramsii	rsa0063086	magnoliopsida	pom	3818e1d3-b6a4-11e8-b408-001a64db2964		p. a. munz
		preservedspecimen		heuchera abramsii	rsa0066369	magnoliopsida	rsa	3818e1d3-b6a4-11e8-b408-001a64db2964		r. f. thorne; c.
		preservedspecimen		heuchera		magnoliopsida	dicots			i.m. johnston
		preservedspecimen		heuchera		magnoliopsida	dicots			r.g. swinney
		preservedspecimen		heuchera		magnoliopsida	dicots			r.g. swinney
		preservedspecimen		heuchera		magnoliopsida	dicots			r.g. swinney
		preservedspecimen		heuchera abramsii	rsa0063160	magnoliopsida	rsa	3818e1d3-b6a4-11e8-b408-001a64db2964		l. e. hoffman
		preservedspecimen		heuchera abramsii	rsa0063087	magnoliopsida	rsa	3818e1d3-b6a4-11e8-b408-001a64db2964		wesley o. grie
		preservedspecimen		heuchera abramsii	csla016633	magnoliopsida		9a7f2f77-3288-4d40-8baa-2f55b8450b7f		r. vandenbur
		preservedspecimen		heuchera abramsii	ucr0048351	magnoliopsida		3818c5b4-b6a4-11e8-b408-001a64db2964		r.g. swinney
		preservedspecimen		heuchera		magnoliopsida	dicots			r.g. swinney
		preservedspecimen		heuchera abramsii	rsa0066355	magnoliopsida	rsa	3818e1d3-b6a4-11e8-b408-001a64db2964		f. r. fosberg; r
		preservedspecimen		heuchera abramsii	dav395249	magnoliopsida		cc54a5fd-2e20-4ea6-8871-b2af436d9de3		r.g. swinney

These columns are found in .dwca files

Opening one of the folders contained in the output directory!

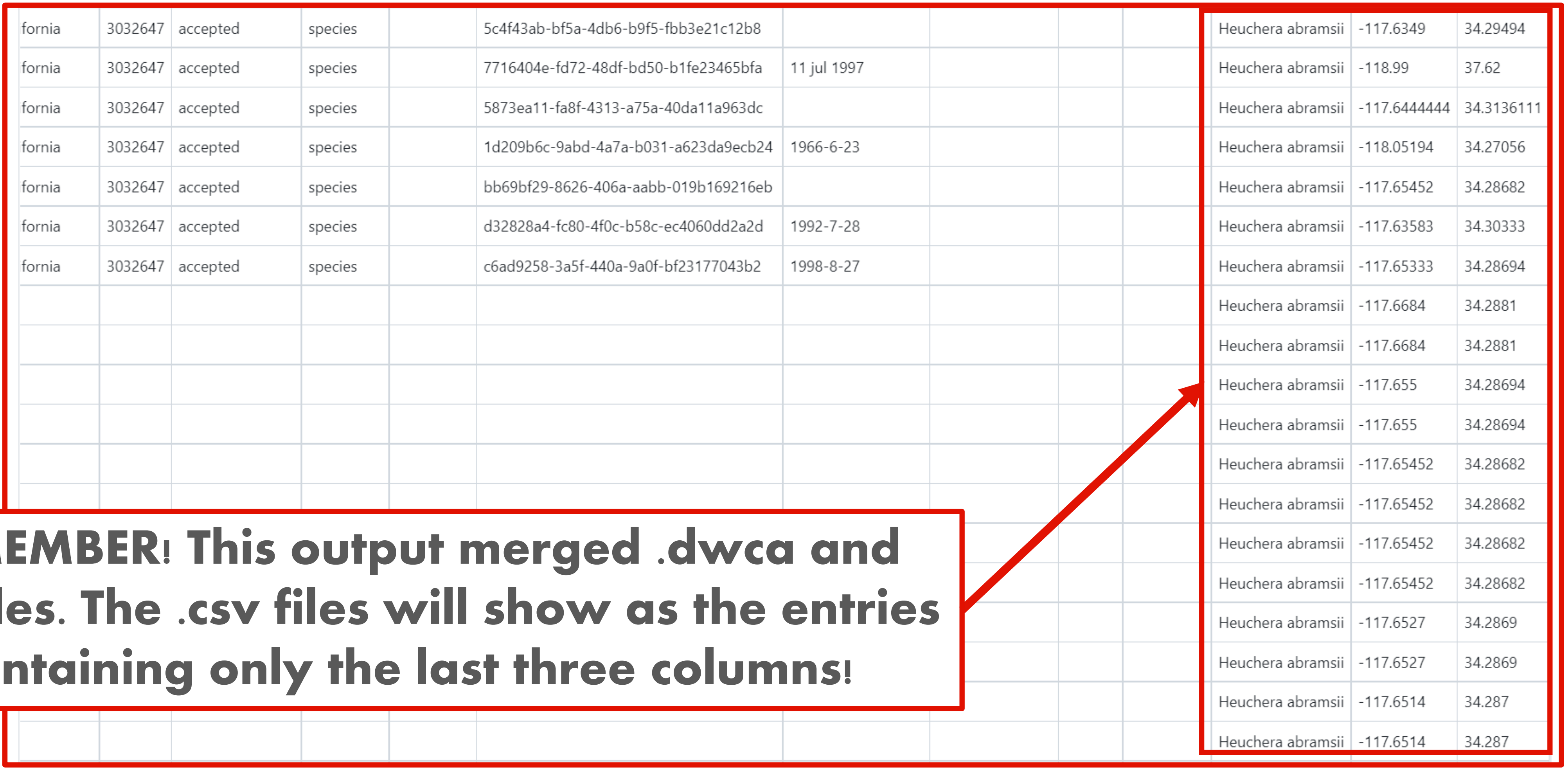
Let's look at the output!

There are many columns in .dwca files, but the ones that will be relevant for now are the last three!



teProvince	taxonID	taxonomicStatus	taxonRank	typeStatus	uuid	verbatimEventDate	verbatimLocality	version	waterBody	species_name	x	y
fornia	3032647	accepted	species		0208288e-4c4a-4d81-aa8f-120ec6327f6b					Heuchera abramsii	-117.6461	34.28921
fornia	3032647	accepted	species		e6f75926-c0ee-42e0-9203-2d6af3f489f5					Heuchera abramsii	-117.65052	34.2878
fornia	3032645	accepted	genus		3a393342-9941-4cf0-9217-875a584f9b78					Heuchera abramsii	-117.64617	34.28917
fornia	3032645	accepted	genus		222f2377-aeaa-4657-8db0-328a4d66cdb0					Heuchera abramsii	-117.655	34.28694
fornia	3032645	accepted	genus		b87e34f8-6cce-4221-8890-8efe6ddc170c					Heuchera abramsii	-117.64444	34.31361
fornia	3032645	accepted	genus		f71d5f6a-fa5b-44c3-aa7f-54eb4e98b891					Heuchera abramsii	-117.64389	34.28861
fornia	3032647	accepted	species		3dced37c-6821-46f8-83a5-dbf4e295b208					Heuchera abramsii	-117.6461	34.28921
fornia	3032647	accepted	species		66daffa9-e2eb-4688-991e-7529688a39e8					Heuchera abramsii	-117.6461	34.28921
fornia	3032647	accepted	species		4bd2eaa8-efb9-4dbe-b762-50ec2a1e7402	9 jul 1967				Heuchera abramsii	-117.9288	34.3493
fornia	3032647	accepted	species		ad455030-61b1-473c-aa5d-154f3ac74c8b	2000-6-22				Heuchera abramsii	-117.5725	34.22833
fornia	3032645	accepted	genus		ee7c56eb-4107-4680-9b12-672709b2679f					Heuchera abramsii	-117.64611	34.28944
fornia	3032647	accepted	species		3e4b21e4-7522-4c6c-8d1b-70baa438f67a					Heuchera abramsii	-117.6461	34.28921
fornia	3032647	accepted	species		14f1dde4-9b83-4790-b298-6e39e2b305e1	7 sep 1998				Heuchera abramsii	-117.64444	34.31361
fornia	3032645	accepted	genus		0358a4cb-924c-443f-b040-589854ba5b6b					Heuchera abramsii	-118.05186	34.2705
fornia	3032647	accepted	species		04450e62-be59-473d-9a5b-222089afe3d2	1918-7-6				Heuchera abramsii	-117.64617	34.28917

Let's look at the output!



fornia	3032647	accepted	species		5c4f43ab-bf5a-4db6-b9f5-fbb3e21c12b8					Heuchera abramsii	-117.6349	34.29494
fornia	3032647	accepted	species		7716404e-fd72-48df-bd50-b1fe23465bfa	11 jul 1997				Heuchera abramsii	-118.99	37.62
fornia	3032647	accepted	species		5873ea11-fa8f-4313-a75a-40da11a963dc					Heuchera abramsii	-117.6444444	34.3136111
fornia	3032647	accepted	species		1d209b6c-9abd-4a7a-b031-a623da9ecb24	1966-6-23				Heuchera abramsii	-118.05194	34.27056
fornia	3032647	accepted	species		bb69bf29-8626-406a-aabb-019b169216eb					Heuchera abramsii	-117.65452	34.28682
fornia	3032647	accepted	species		d32828a4-fc80-4f0c-b58c-ec4060dd2a2d	1992-7-28				Heuchera abramsii	-117.63583	34.30333
fornia	3032647	accepted	species		c6ad9258-3a5f-440a-9a0f-bf23177043b2	1998-8-27				Heuchera abramsii	-117.65333	34.28694
										Heuchera abramsii	-117.6684	34.2881
										Heuchera abramsii	-117.6684	34.2881
										Heuchera abramsii	-117.655	34.28694
										Heuchera abramsii	-117.655	34.28694
										Heuchera abramsii	-117.65452	34.28682
										Heuchera abramsii	-117.65452	34.28682
										Heuchera abramsii	-117.65452	34.28682
										Heuchera abramsii	-117.65452	34.28682
										Heuchera abramsii	-117.6527	34.2869
										Heuchera abramsii	-117.6527	34.2869
										Heuchera abramsii	-117.6514	34.287
										Heuchera abramsii	-117.6514	34.287

REMEMBER! This output merged .dwca and .csv files. The .csv files will show as the entries containing only the last three columns!

Session Summary, Q&A and Discussion

- ✓ There are multiple ways and repositories to access occurrence data.
- ✓ Phylogenetic trees can be obtained from the Open Tree of Life, as well as other sources.
- ✓ BiotaPhy tools enable automated data downloads of occurrences and trees.
- ✓ BiotaPhy tools split large data sets into species-specific data sets for SDM and other uses.

Session Summary, Q&A and Discussion

Any questions??

Please use the chat to ask your questions!