

1 rgbif: R client for working with GBIF species occurrence data

2 Scott Chamberlain^{*,a}

3 ^a*University of California, Berkeley, CA, USA*

4 **Abstract**

- 5 1. xxx
- 6 2. xxx
- 7 3. xxx
- 8 4. xxxx

9 **Introduction**

10 Perhaps the most fundamental element in many fields of ecology is the individual. How many individuals
11 of each species in a given location forms the basis for many sub-fields of ecology and evolution. Some
12 research questions necessitate collecting new data, while others can easily take advantage of existing
13 data. In fact, some ecology fields are built largely on existing data, e.g., macro-ecology (Brown, 1995;
14 Beck et al., 2012).

15 Data on individuals, including which species, and where they're found, can be used for a large number
16 of research questions. In addition, the pool of questions we can answer becomes much larger with more
17 and better data. In addition to wide utility, this data is important for conservation. Biodiversity loss is
18 one of the greatest challenges of our time (Pimm et al., 2014). Some have called this the sixth great
19 mass extinction (Ceballos et al., 2015). Given this challenge there is a great need for data on specimen
20 records, whether collected from live sightings in the field or specimens in museums.

21 There are many online services that collect and maintain specimen records. However, Global Biodiversity
22 Information Facility (hereafter, GBIF, <http://www.gbif.org/>) is the largest collection of biodiversity
23 records globally, currently with 580 million records, 1.6 million taxa, 15,000 datasets from 770 publishers
24 (figures collected on 2015-10-04). Many large biodiversity warehouses such as iNaturalist, VertNet, and
25 USGS's BISON all feed into GBIF.

26 Herein, we describe a library (rgbif (Chamberlain et al.)) for working with GBIF data in the R
27 programming environment (R Core Team, 2014). R is an extremely widely used language in academia,

*Corresponding author
Email address: `scott(at)ropensci.org` (Scott Chamberlain)

28 and in non-profit and private sectors. Importantly, R makes it easy to do all of the steps of the research
29 process, including data management, data manipulation and cleaning, statistics, and vizualization.
30 Thus, an R client for getting GBIF data is a powerful tool for reproducible research.

31 **The rgbif package**

32 The `rgbif` package is completely written in R, uses an [MIT license](#) to maximize use everywhere. `rgbif`
33 is developed publicly on GitHub at <https://github.com/ropensci/rgbif>, where development versions of
34 the package can be installed, and bugs and feature requests reported. Stable versions of `rgbif` can be
35 installed from [CRAN](#), the distribution network for R packages. `rgbif` is part of the rOpenSci project,
36 a developer network making R software to facilitate reproducible research.

37 *Package interface*

38 `rgbif` is designed following the [GBIF Application Programming Interface](#), or API. The GBIF API has
39 four major components: registry, species names, occurrence data, and maps. We ignore maps in `rgbif`
40 as it is concerned with generating maps for web applications. `rgbif` has a suite of functions dealing
41 with each of registry, species names, and occurrence data - we'll go through each in turn describing
42 design and example usage.

43 *Registry*

44 The GBIF registry API services are spread across four sets of functions:

- 45 • Datasets
- 46 • Installations
- 47 • Networks
- 48 • Nodes
- 49 • Organizations

50 *Datasets*

51 Search for datasets

```
res <- dataset_search(query = "oregon")
res$data$datasetTitle[1:10]
```

```
52 [1] "SDNHM Birds Collection"
53 [2] "CM Birds Collection"
54 [3] "condoncollection"
55 [4] "Taxonomy in Flux Checklist"
56 [5] "Wool carder bees of the genus Anthidium in the Western Hemisphere"
57 [6] "Bryophyte Collection - University of Washington Herbarium (WTU)"
58 [7] "University of British Columbia Herbarium (UBC) - Bryophytes Collection"
59 [8] "UWFC Ichthyology Collection"
60 [9] "Lichen Collection - University of Washington Herbarium (WTU)"
61 [10] "UWBM Mammalogy Collection"
```

62 Get dataset metrics

```
res <- dataset_metrics(uuid='66dd0960-2d7d-46ee-a491-87b9adcfe7b1')
df <- data.frame(rank = names(res$countByRank),
                 count = unname(unlist(res$countByRank)))
knitr::kable(df)
```

rank	count
SPECIES	52452
GENUS	12930
VARIETY	4806
SUBSPECIES	4440
SERIES	1079
TRIBE	844
FAMILY	509
SUBTRIBE	327
SUBFAMILY	303
SUBGENUS	241

rank	count
FORM	239
SECTION	82
SUBVARIETY	4
KINGDOM	1

63 *Networks, nodes, and installations*

64 Here, we search for the first give GBIF networks, returning just the key and title fields.

```
networks(limit=10)$data$title
```

```
65 [1] "GBIF Backbone Sources"
66 [2] "Canadensys"
67 [3] "Southwest Collections of Arthropods Network (SCAN)"
68 [4] "VertNet"
69 [5] "Dryad"
70 [6] "GBIF Network"
71 [7] "The Knowledge Network for Biocomplexity (KNB) "
72 [8] "Online Zoological Collections of Australian Museums (OZCAM)"
73 [9] "Catalogue of Life"
74 [10] "Ocean Biogeographic Information System (OBIS)"
```

75 *Species*

76 *Occurrences*

77 GBIF provides two ways to get occurrence data: through the `/occurrence/search` route (see
78 `occ_search()`), or via the `/occurrence/download` route (many functions, see below). `occ_search()` is
79 the main funtion for the search route, and is more appropriate for smaller data, while `occ_download*()`
80 functions are more appropriate for larger data requests.

81 Large is of course a subjective term. When you hit a “large dataset” will depend primarily on the size
82 of the your data request. GBIF imposes for any given search a limit of 200,000 records in the search

83 service, after which point you can't download any more records for that search. However, you can
84 download more records for different searches.

85 We think the search service is still quite useful for many people even given the 200,000 limit. For those
86 that need more data, we have created a similar interface in the `download_*`() functions, that should be
87 easy to use. Users should take note that using the download service has a few extra steps to get data
88 into R, but is straight-forward.

89 *Download API*

90 The download API syntax is similar to the occurrence search API in that the same parameters are used,
91 but the way in which the query is defined is different. For example, in the download API you can do
92 greater than searches (i.e., `latitude > 50`), whereas you can not do that in the occurrence search API.
93 Thus, we can't make the query interface exactly the same for both search and download functions.

94 Using the download service can be as few as three steps: 1) Request data via a search; 2) Download
95 data; 3) Import data into R.

96 Request data download given a query. Here, we search for the taxon key 3119195, which is the key for
97 *Helianthus annuus* (<http://www.gbif.org/species/3119195>).

```
occ_download('taxonKey = 3119195')
```

98 <<gbif download>>

99 Username: sckott

100 E-mail: myrmecocystus@gmail.com

101 Download key: 0004098-150922153815467

102 You can check on when the download is ready using the functions `occ_download_list()` and
103 `occ_download_meta()`. When it's ready use `occ_download_get()` to download the dataset to your
104 computer.

```
(res <- occ_download_get("0000840-150615163101818", overwrite = TRUE))
```

105 <<gbif downloaded get>>

106 Path: ./0000840-150615163101818.zip

107 File size: 3.19 MB

108 What's printed out above is a very brief summary of what was downloaded, the path to the file, and its
109 size (in human readable form).

110 Next, read the data in to R using the function `occ_download_import()`.

```
library("dplyr")
dat <- occ_download_import(res)
dat %>%
  select(gbifID, decimalLatitude, decimalLongitude)
```

	gbifID	decimalLatitude	decimalLongitude
112	1 657590544	NA	NA
113	2 657679551	NA	NA
114	3 657791316	37.70805	-118.4162
115	4 658180562	NA	NA
116	5 441881672	NA	NA
117	6 911596181	NA	NA
118	7 56454601	NA	NA
119	8 657848913	NA	NA
120	9 658187373	NA	NA
121	10 658279212	38.95917	-106.9892
122

123 *Downloaded data format.* The downloaded dataset from GBIF is actually a Darwin Core Archive
124 (DwC-A), an internationally recognized biodiversity informatics standard (<http://rs.tdwg.org/dwc/>).
125 The DwC-A downloaded is a compressed folder with a number of files, including metadata, citations for
126 each of the datasets included in the download, and the data itself, in separate files for each dataset as
127 well as one single `.txt` file. In `occ_download_import()`, we simply fetch data from the `.txt` file. If you
128 want to dig into the metadata, citations, etc., it is easily accessible from the folder on your computer.

129 *Search API*

130 The search API follows the GBIF API and is broken down into the following functions:

- 131 • `occ_count()`

```
132 • occ_search()
133 • occ_get()
134 • occ_metadata()
```

135 The main search work-horse is `occ_search()`. This function allows very flexible search definitions. In
136 addition, this function does paging internally, making it such that the user does not have worry about
137 the 300 records per request limit - but of course we can't go over the 200,000 maximum limit.

138 ...

139 *Cleaning data.* GBIF provides optional data issues with each occurrence record. These issues fall into
140 many different pre-defined classes, covering issues with taxonomic names, geographic data, and more
141 (see `occ_issues_lookup()` to find out more information on GBIF issues; and the same data on [GBIF's](#)
142 [development site](#)).

143 `occ_issues()` provides a way to easily filter data downloaded via `occ_search()` based on GBIF issues.

```
out <- occ_search(issue='DEPTH_UNLIKELY', limit = 500)
NROW(out)
```

```
144 [1] 4
```

```
out %>% occ_issues(-cudc) %>% .$data %>% NROW
```

```
145 [1] 2
```

146 *Use cases*

147 ...

148 Conclusions and future directions

149 Acknowledgements

150 This project was supported in part by the Alfred P Sloan Foundation (Grant 2013-6-22).

151 *Data Accessibility*

152 All scripts and data used in this paper can be found in the permanent data archive Zenodo under
153 the digital object identifier (DOI). This DOI corresponds to a snapshot of the GitHub repository at
154 github.com/sckott/msrgbif. Software can be found at github.com/ropensci/rgbif, under the open and
155 permissive MIT license.

156 **References**

- 157 Beck J., Ballesteros-Mejia L., Buchmann CM., Dengler J., Fritz SA., Gruber B., Hof C., Jansen
158 F., Knapp S., Kreft H., Schneider A-K., Winter M., Dormann CF. 2012. Whats on the horizon for
159 macroecology? *Ecography* 35:673–683.
- 160 Brown JH. 1995. *Macroecology*. University of Chicago Press.
- 161 Ceballos G., Ehrlich PR., Barnosky AD., Garcia A., Pringle RM., Palmer TM. 2015. Accelerated
162 modern human-induced species losses: Entering the sixth mass extinction. *Science Advances* 1:e1400253–
163 e1400253.
- 164 Chamberlain S., Ram K., Barve V., Mcglinn D. *Rgbif: Interface to the global 'biodiversity' information*
165 *facility 'aPI'*.
- 166 Pimm SL., Jenkins CN., Abell R., Brooks TM., Gittleman JL., Joppa LN., Raven PH., Roberts CM.,
167 Sexton JO. 2014. The biodiversity of species and their rates of extinction, distribution, and protection.
168 *Science* 344:1246752–1246752.
- 169 R Core Team. 2014. *R: A language and environment for statistical computing*. Vienna, Austria: R
170 Foundation for Statistical Computing.