rgbif: a package for working with species occurrence data in R

Scott Chamberlain[*,a]

[a]*University of California, Berkeley, CA, USA*

Abstract

1. xxx

2. xxx

3. xxx

4. xxxx

Introduction

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

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

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

Herein, we describe a library (rgbif (Chamberlain *et al.*)) for working with GBIF data in the R 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)

October 4, 2015

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.

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.

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.

Registry

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

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

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 `occ_search`),

78 or via the /occurrence/download route (many functions, see below). `occ_search` is more appropriate

79 for smaller data, while occ_download*() functions are more appropriate for larger data requests.

80 *Download API*

81 The download API syntax is similar to the occurrence search API in that the same parameters are used,

82 but the syntas is quite different. For example, in the download API you can do greater than searches,

83 whereas you can not do that in the occurrence search API.

```
occ_download_get("0000066-140928181241064")
```

```
84  <<gbif downloaded get>>
85    Path: ./0000066-140928181241064.zip
86    File size: 0.14 MB
```

Search API

The search API is very similar to the download API, but is meant for smaller data acquisition jobs.

Conclusions and future directions

Acknowledgements

Data Accessibility

All software, scripts and data used in this paper can be found in the permanent data archive Zenodo under the digital object identifier (DOI). This DOI corresponds to a snapshot of the GitHub repository at github.com/sckott/msrgbif.

References

Beck, J., Ballesteros-Mejia, L., Buchmann, C.M., Dengler, J., Fritz, S.A., Gruber, B., Hof, C., Jansen, F., Knapp, S., Kreft, H., Schneider, A.-K., Winter, M. & Dormann, C.F. (2012). Whats on the horizon for macroecology? *Ecography*, **35**, 673–683. Retrieved from http://dx.doi.org/10.1111/j.1600-0587.2012.07364.x

Brown, J.H. (1995). *Macroecology.* University of Chicago Press.

Ceballos, G., Ehrlich, P.R., Barnosky, A.D., Garcia, A., Pringle, R.M. & Palmer, T.M. (2015). Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances*, **1**, e1400253–e1400253. Retrieved from http://dx.doi.org/10.1126/sciadv.1400253

Chamberlain, S., Ram, K., Barve, V. & Mcglinn, D. *Rgbif: Interface to the global 'biodiversity' information facility 'aPI'.* Retrieved from https://github.com/ropensci/rgbif

Pimm, S.L., Jenkins, C.N., Abell, R., Brooks, T.M., Gittleman, J.L., Joppa, L.N., Raven, P.H., Roberts, C.M. & Sexton, J.O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, **344**, 1246752–1246752. Retrieved from http://dx.doi.org/10.1126/science.1246752

R Core Team. (2014). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/