Species occurrence data (SPOCC), version 0.08

Introduction

The rOpenSci projects aims to provide programmatic access to scientific data repositories on the web. A vast majority of the packages in our current suite retrieve some form of biodiversity or taxonomic data. Since several of these datasets have been georeferenced, it provides numerous opportunities for visualizing species distributions, building species distribution maps, and for using it analyses such as species distribution models. In an effort to streamline access to these data, we have developed a package called Spocc, which provides a unified API to all the biodiversity sources that we provide. The obvious advantage is that a user can interact with a common API and not worry about the nuances in syntax that differ between packages. As more data sources come online, users can access even more data without significant changes to their code. However, it is important to note that spocc will never replicate the full functionality that exists within specific packages. Therefore users with a strong interest in one of the specific data sources listed below would benefit from familiarising themselves with the inner working of the appropriate packages.

Data Sources

`SPOCC` currently interfaces with five major biodiversity repositories. Many of these packages have been part of the rOpenSci su

1. Global Biodiversity Information Facility (`rgbif`)
   Gbif is a government funded open data repository with several partner organizations with the express goal of providing access to data on Earth's biodiversity. The data are made available by a network of member nodes, coordinating information from various participant organizations and government agencies.

2. Berkeley Ecoengine (`ecoengine`)
   The ecoengine is an open API built by the Berkeley Initiative for Global Change Biology. The repository provides access to over 3 million specimens from various Berkeley natural history museums. These data span more than a century and provide access to georeferenced specimens, species checklists, photographs, vegetation surveys and resurveys and a variety of measurements from environmental sensors located at reserves across University of California's natural reserve system.

3. **iNaturalist** (`inat`)
   iNaturalist provides access to crowd sourced citizen science data on species observations.

4. Vertnet (`vertnet`) Similar to gbif, ecoengine, and bison (see below), Vernet provides access to more than 80 million vertebrate records spanning a large number of institutions and museums primarly covering four major disciplines (mammology, herpetology, ornithology, and icthyology).

5. Biodiversity Information Serving Our Nation (`rbison`)
   Built by the US Geological Survey's core science analytic team, BISON is a portal that provides access to species occurrence data from several participating institutions.

6. ebird (`rebird`)
   ebird is a database developed and maintained by the Cornell Lab of Ornithology and the National Audubon Society. It provides real-time access to checklist data, data on bird abundance and distribution, and communtiy reports from birders.

Notes: It's important to keep in mind that several data providers interface with many of the above mentioned repositoires. This means that occurence data obtained from BISON may be duplicates of data that are also available through GBIF. We do not have a way to resolve these duplicates or overlaps at this time but it is an issue we are hoping to resolve in future versions of the package.

Data retrival

The most significant function in spocc is the `occ` (short for occurrence) function. `occ` takes a query, often a species name, and searches across all data sources specified in the `from` argument. For example, one can search for all occurrences of Sharp-shinned Hawks (*Accipiter striatus*) from the gbif database with the following R call.

```
library(spocc)
df <- occ(query = "Accipiter striatus", from = "gbif")
```

```
## Loading required package: rjson
```

```
df
```

```
## Summary of results - occurrences found for:
##  gbif  : 25 records across 1 species
##  bison :  0 records across 1 species
##  inat  :  0 records across 1 species
##  ebird :  0 records across 1 species
##  ecoengine :  0 records across 1 species
```

```
head(df$gbif$data[[1]])
```

```
##                  name       key longitude latitude prov
## 1 Accipiter striatus 768992325    -76.10    4.724 gbif
## 2 Accipiter striatus 773408845    -97.32   32.821 gbif
## 3 Accipiter striatus 773414146   -122.27   37.771 gbif
## 4 Accipiter striatus 859267562   -108.34   36.732 gbif
## 5 Accipiter striatus 859267548   -108.34   36.732 gbif
## 6 Accipiter striatus 859267717   -108.34   36.732 gbif
```

The data returned are part of a `S3` class called `occdat`. This class has slots for the various data sources described above. One can easily switch the source by changing the `from` in the function call above.

Visualization routines

Interactive maps

Leaflet JS
Leaflet JS is an open source mapping library that can leverage various layers from multiple sources. Using the `leafletR` library, it's possible to generate a local `geoJSON` file and a html file of species distribution maps. The folder can easily be moved to a web server and served widely without any additional coding.

It's also possible to render similar maps iwth Mapbox by committing just the geoJSON file to GitHub or posting it as a gist. All the remaining fields will become part of a table inside a tooltip, providing a extremely quick and easy way to serve up interactive maps. This is especially useful when a user does not have access to web hosting.

Shiny map

Some text...

Static maps

Describe the static ggmap.

- Should we still describe the other (rCharts) stuff even though those functions wont be in the intial release?

Use-cases

Some text...

Upcoming features

Some text...

References

Some text...