rOpenSci Data Packages

Scott Chamberlain

Why Data Packages?

- Reduce duplicated effort by each researcher
- One best way to get data XYZ
- Reduced user error
- Allow researchers to focus on the science

Data Packages: Caveats

- User base (# of people) for data pkgs small relative to utilities
  - ... attracts fewer contributors
  - ... & all the carry on effects of above
- Pkg can get out of sync with data source's API
- Risk leaving out metadata/context

rOpenSci Data Packages

- Biological occurrence data
- Taxonomy data
- Climate data
- Geospatial data/tools

Occurrence Data Packages

- spocc - Biodiversity data toolbelt
- rgbif - GBIF data (avail. in spocc)
- ecoengine - Berekeley Ecoengine client (avail. in spocc)
- rinat - iNaturalist client (avail. in spocc)
- rbison - USGS BISON client (avail. in spocc)
- rebird - eBird data via their API (avail. in spocc)
- auk - eBird bulk data
- rvertnet - VertNet data (avail. in spocc)
- rfishbase - Fishbase.org data
- finch - xx

rgbif usage: Checklist recipe

TrIAS Project - a template for standardizing species checklist data to Darwin Core using R

```
├── README.md                : Description of this repository
├── LICENSE                  : Repository license
├── checklist-recipe.Rproj : RStudio project file
├── .gitignore               : Files and directories to be ignored by git
|
├── data
|   ├── raw                  : Source data, input for mapping script
|   └── processed            : Darwin Core output of mapping script GENERATED
|
├── docs                     : Repository website GENERATED
|
└── src
    ├── dwc_mapping.Rmd      : Darwin Core mapping script, core functionality of this r
    ├── _site.yml            : Settings to build website in /docs
    └── index.Rmd            : Template for website homepage
```

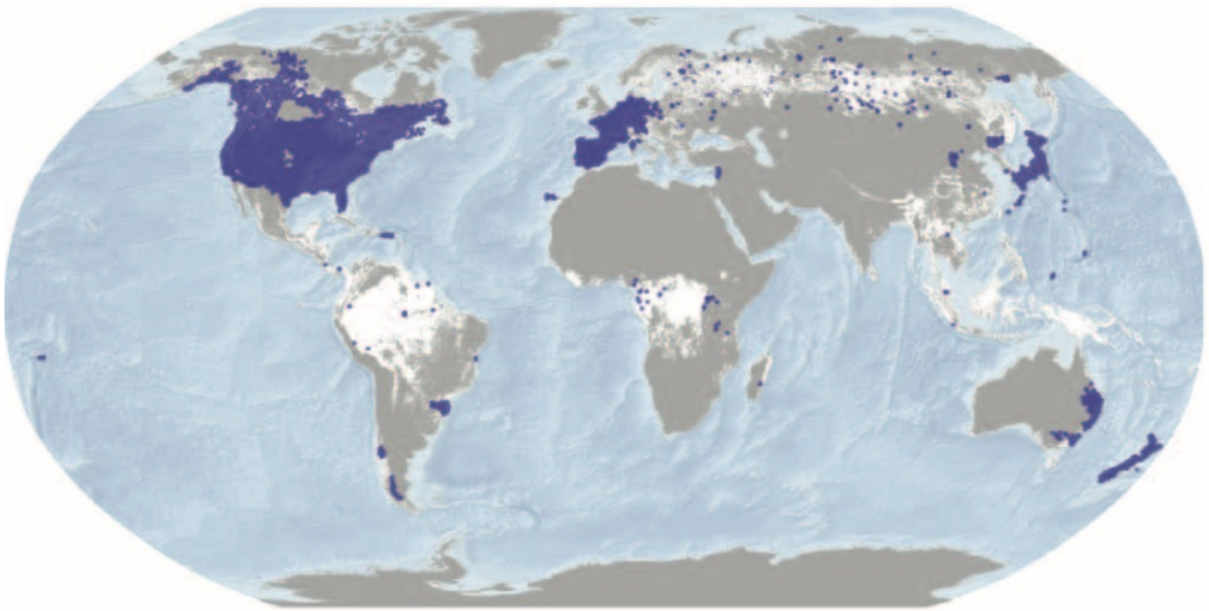
Read post on the rOpenSci blog

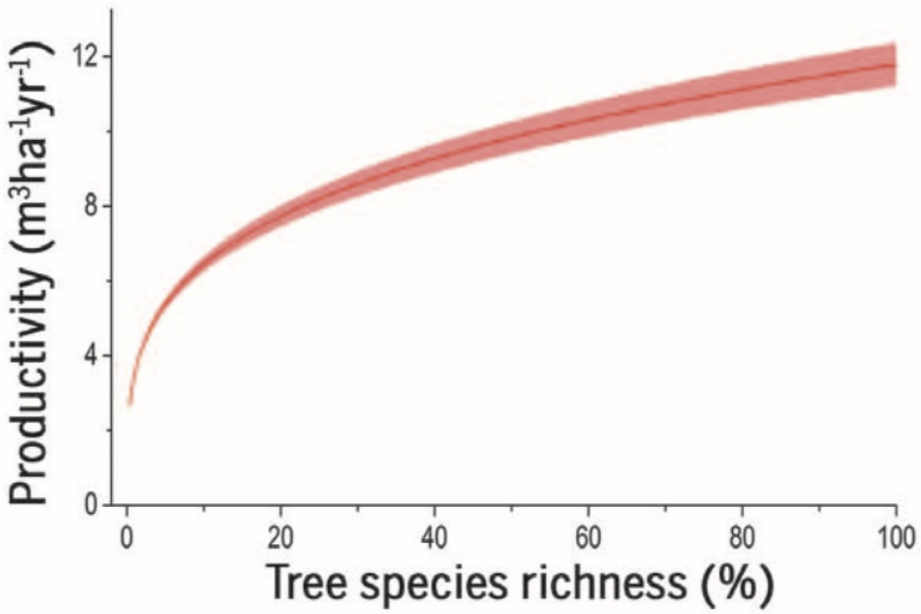
Taxonomy Packages

- taxa - Taxonomic toolbelt
- taxize - Taxonomic toolbelt
- taxizedb - Taxonomic toolbelt - local database backed

taxize usage

Liang, J., et al. (2016) -- Positive biodiversity-productivity relationship predominant in global forests





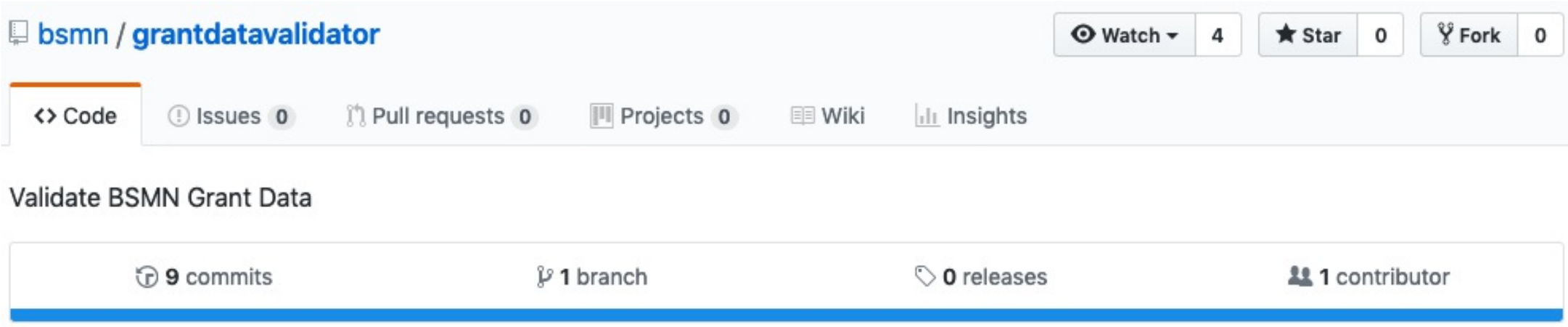
> *there were ... 8,737 species ... We verified all ... names against 60 taxonomic data-bases, including NCBI, GRIN Taxonomy for Plants, Tropicos–Missouri Botanical Garden, and the International Plant Names Index, using the 'taxize' package in R*

Utility Packages

- jqr - xx
- jsonld - xx
- rerddap - xx
- rdflib - xx
- assertr - xx

assertr usage: eg

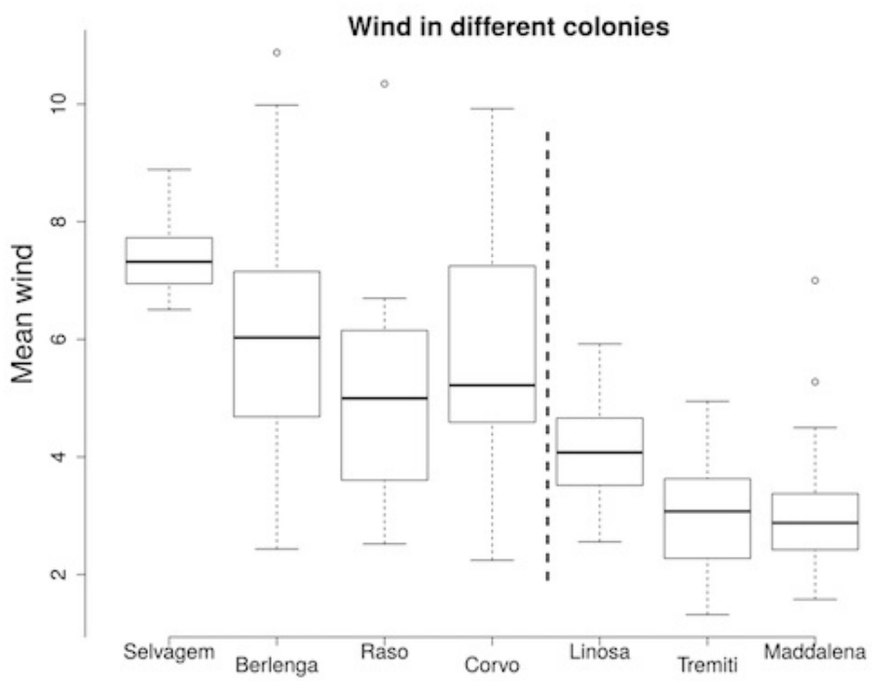
Brain Somatic Mosaicism Network - Validate BSMN Grant Data




```
data %>%
   assertr::chain_start() %>%
   assertr::verify(nrow(data) == 3) %>%
   assertr::verify(assertr::is_uniq(nda_short_name)) %>%
   assertr::verify(assertr::not_na(grant)) %>%
   assertr::verify(dplyr::n_distinct(grant) == 1) %>%
   assertr::verify(nda_short_name %in% expected_nda_short_names) %>%
   assertr::chain_end() %>%
   tibble::as_tibble()
```

rerddap usage: eg

Abolaffio, J., et al. (2018) -- Olfactory-cued navigation in shearwaters: linking movement patterns to mechanisms



> *wind data were downloaded from the NOAA28 web site from the rerdapp package for R*

Data integration: Steps

- Start with a species list
- Clean names with taxize
- Get occurrence data with rgbif
- Clean occurrence data with scrubr and CoordinateCleaner
- assertr to check data
- Map with mapr

Data integration: code

```r
#spp <- names_list("species")
spp <- read.csv("spp_list.txt", header = TRUE, stringsAsFactors = FALSE)$bad
spp2 <- taxize::gnr_resolve(spp, data_source_ids=11, canonical=TRUE)$matched
dat <- rgbif::occ_data(scientificName = spp2, limit = 100)
dat
```

```
## Occ. found [Cyperu.. (16), Lagose.. (164), Pandan.. (27), Gentia.. (11),
##      Spheno.. (560), Sphagn.. (159), Trades.. (117), Cloezi.. (53),
##      Saxifr.. (1686), Baccha.. (1147)]
## Occ. returned [Cyperu.. (16), Lagose.. (100), Pandan.. (27), Gentia..
##      (11), Spheno.. (100), Sphagn.. (100), Trades.. (100), Cloezi.. (53),
##      Saxifr.. (100), Baccha.. (100)]
## Args [limit=100, offset=0, scientificName=Cyperus kappleri,Lagoseris
##      sancta,Pandanus polyglossus,Gentianella
##      scarlatinostriata,Sphenolobopsis ]
## 10 requests; First 10 rows of data from Cyperus kappleri
##
## # A tibble: 16 x 93
##    name     key decimalLatitude decimalLongitude issues datasetKey
##    <chr>  <int>           <dbl>            <dbl> <chr>  <chr>
##  1 Cype… 1.26e9            6.93            -71.2 gass84 7bd65a7a-…
##  2 Cype… 1.23e9          -11.3             -67.7 gass84 90c853e6-…
##  3 Cype… 1.26e9          -13.6             -60.8 gass84 7bd65a7a-…
```