rOpenSci Data Packages

Scott Chamberlain

@sckottie
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 occurrences
- Taxonomy
- Data utilities

find more: ropensci.org/packages
Occurrence Data Packages

- **spocc** - Biodiversity data toolbelt
- **rgbif** - GBIF data
- **ecoengine** - Berkeley Ecoengine client
- **rinat** - iNaturalist client
- **rbison** - USGS BISON client
- **rebird** - eBird data via their API
- **auk** - eBird bulk data
- **rvertnet** - VertNet data
- **rfishbase** - Fishbase.org data
- **finch** - Handle Darwin Core
Data package patterns

Or at least patterns we strive towards ...

- HTTP requests
- Cache downloaded data for FTP/similar files
- Return data.frame's: facilitates downstream processing
- Make it easy to cite data providers
- Incorporate metadata
rgbif

Access > 1 billion occurrence records from GBIF

ropensci/rgbif

- Search for species
- Download occurrence data
- Clean occurrence data
- Make maps
rgbif usage


... records were obtained for each species ... using ... rgbif ... Quality controls were applied to ensure that occurrence records were correctly taxonomically identified and accurately georeferenced. In particular, records with the following issues were excluded ... invalid coordinates or geodetic datum; failed or suspicious coordinate reprojection; ... species name with no match ... 

~70 more e.g.'s at https://github.com/ropensci/roapi/blob/master/data/citations.csv
rgbif usage: Checklist recipe

TrIAS Project - 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
    └── _site.yml          : Settings to build website in /docs
        └── index.Rmd     : Template for website homepage

rgbif
Taxonomy Packages

- **taxa** - Taxonomic R classes
- **taxize** - Taxonomic toolbelt (remote data)
  - **taxize book**
- **taxizedb** - Leverage dumps of taxonomic database locally
- **ritis** - USGS's ITIS
- **rotl** - Open Tree of Life
- **rentrez** - NCBI ENTREZ taxonomy database
- **worrms** - WORMS marine taxonomy
- **wikitaxa** - Taxonomy data on Wikipedia
- **zbank** - ZooBank
- **rgbif** - GBIF taxonomy data
taxize

Access taxonomic data from > 20 sources

ropensci/taxize

- Resolve misspelled/etc. names
- Search for names
- Taxonomic classifications
- Fetch all taxa up- and down-stream
- Fetch taxonomic synonyms
- Common names to taxonomic and vice versa
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

~90 more e.g.'s at https://github.com/ropensci/roapi/blob/master/data/citations.csv
Utility Packages

- jqr - R client for jq, the JSON processor
- jsonld - JSON linked data
- rerddap - ERDDAP server client
- rdflib - RDF pkg, wrapped around Redland
- assertr - Assertions for analysis pipelines
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


wind data were downloaded from the NOAA web site from the rerddapp package for R
Data integration: Steps

- Start with a species list
- Clean names with `taxize`
- Get occurrence data with `rgbif`
- Clean occurrence data w/ `rgbif`, `scrubr` or `CoordinateCleaner`
- `assertr` to check data
- Map with `mapr`
Data integration: code

```r
# read in species list
spp <- read.csv("spp_list.txt", header = TRUE,
                stringsAsFactors = FALSE)$bad
# resolve names: fix misspellings
spp2 <- taxize::gnr_resolve(spp, data_source_ids = 11,
                            canonical = TRUE)$matched_name2
# fetch GBIF occurrence data
dat <- rgbif::occ_data(scientificName = spp2, limit = 300)
# remove data with issues: COUNTRY_MISMATCH & COORDINATE_ROUNDED
dat <- rgbif::occ_issues(dat, -cum, -cdround)
# make a single data.frame
dat <- dplyr::bind_rows(lapply(dat, "[[", "data"))
# remove records with incomplete lat/lon data
dat <- scrubr::coord_incomplete(dat)
```

Data: `spp_list.txt`
Data integration: code

Test assertions about the data

```r
library(magrittr)
dat %>%
  assertr::chain_start() %>%
  # does it have more than 100 rows?
  assertr::verify(NROW(dat) > 100) %>%
  # is the key for the occurrence record unique?
  assertr::verify(assertr::is_uniq(key)) %>%
  # are there any NA's in lat/lon?
  assertr::verify(assertr::not_na(decimalLatitude)) %>%
  assertr::verify(assertr::not_na(decimalLongitude)) %>%
  assertr::chain_end() %>%
tibble::as_tibble()
```
Data integration: code

```r
mapr::map_leaflet(dat, lon = "decimalLongitude", lat = "decimalLatitude")
```
Thanks!

Scott Chamberlain

Twitter: sckottie
Slides: scotttalks.info/dataone19

Karthik Ram

Twitter: _inundata

rOpenSci: ropensci.org
A guide to making your data analysis more reproducible

(rough and incomplete)

(and a bunch of other work you do)

Karthik Ram • rOpenSci // UC Berkeley
1. Vorübergehend offene Frage
2. CSV-Datei, Dattenspeicherung, Buch, Dose, [...]
3. Lesen
4. Werkzeugkasten
5. Leere Wange
6. X
"...We introduce the concept of a *compendium* as both a *container* for the different elements that make up the *document* and its computations (i.e. *text, code, data, ...*), and as a means for *distributing, managing* and *updating* the collection.

Gentleman and Temple Lang, 2004
## Research compendium principles

<table>
<thead>
<tr>
<th><img src="image1.png" alt="Icon" /></th>
<th>Stick with the conventions of your peers</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image2.png" alt="Icon" /></td>
<td>Keep data, methods and outputs separate</td>
</tr>
<tr>
<td><img src="image3.png" alt="Icon" /></td>
<td>Specify your computational environment as clearly as you can</td>
</tr>
</tbody>
</table>
Key components you'll need for sharing a compendium
Key components you’ll need for sharing a compendium

License + VCS + Metadata + Archive
The R package structure is great way to organize and share a compendium!
Package: glue
Title: Interpreted String Literals
Version: 1.3.0.9000
Authors@R: person("Jim", "Hester", email = "james.f.hester@gmail.com", role = c("aut", "cre"))
Depends:
  R (>= 3.1)
Imports:
  methods
Suggests:
  testthat,
  (and many more)
License: MIT + file LICENSE
Encoding: UTF-8
LazyData: true
RoxygenNote: 6.0.1
Roxygen: list(markdown = TRUE)
URL: https://github.com/tidyverse/glue
BugReports: https://github.com/tidyverse/glue/issues
VignetteBuilder: knitr
ByteCompile: true
Type: Compendium
Package: pomdpintro
Version: 0.1.0
Depends: nimble, tidyverse, sarsop, MDPtoolbox
Suggests: extrafont, hrbrthemes, Cairo, ggthemes
Remotes: boettiger-lab/sarsop
Packaging your analysis as a compendium gives you access to powerful developer tools
Small compendia

Marwick et al 2017
Medium compendia

Marwick et al 2017
Large/complex compendia

Marwick et al 2017
codemetadata
Add software metadata to a repository

github.com/ropensci/codemetadata
Software testing is important, but, in part because i
Data (Small → Medium)

Computing environment

Workflows
1. Data

How does one manage small to medium data in the context of a research compendium?
Small data

Put small data inside packages, especially if you ship a methods package with your analysis.

CRAN = < 5 mb.

37% of the 13K packages on CRAN have some form of data.
piggyback

Attach large [data] files to Github repositories

github.com/ropensci/piggyback
Leveraging Github releases to share medium sized files

github.com/ropensci/piggyback

```
 pb_new_release("user/repo", "v0.0.5")
 pb_upload("datasets.tsv.xz", "user/repo")
 # Access them in your scripts with
 # pb_download
```
Compendium release v-1.0

karthik released this 3 minutes ago

Assets 3

mtcars.tsv.xz

Source code (zip)

Source code (tar.gz)

Initial release of project datasets
Medium data

github.com/ropensci/arkdb
2. Isolate your computing environment

It's important to isolate the computing environment so that changes in software dependencies don't break your analysis.
Adding a Dockerfile to your compendium

Many ways to write a Dockerfile for your R project

- o2r/containerit
- jupyter/repo2docker
Binder
mybinder.org

Binder is an open source project that is designed to make it really easy to share analyses that are in notebooks.
Resolving the measurement uncertainty paradox in ecological management

- Authors: Milad Memarzadeh, Carl Boettiger

Contents

- **Manuscript**: R Markdown source document for manuscript. Includes code to reproduce for figures from tables generated by the analysis.
- **Appendix**: R Markdown source documents for both appendices, containing all necessary R code to generate all
Loading repository: karthik/binder-test-fastest/master

---
114a4cd0b227
---
Running in 2dad32d8d3e2
Resolving github.com (github.com)... 192.30.253.112, 192.30.253.113
Connecting to github.com (github.com)[192.30.253.112]:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/karthik/binder-test-fastest/master/DESCRIPTION [following]
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.0.133, 151.101.64.133, 151.101.128.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)[151.101.0.133]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 773 [text/plain]
Saving to: ‘DESCRIPTION.1’

OK

2019-01-13 02:04:34 (22.7 MB/s) - ‘DESCRIPTION.1’ saved [773/773]
Git + Docker + RStudio
Setting up Binder

Branch: master

karthik Updated README

- README.md
- code.R
- install.R
- runtime.txt

r-2018-12-20
Setting up Binder

```r
install.packages("ggplot2")
```
Build and launch a repository

GitHub repository name or URL

Git branch, tag, or commit

Path to a notebook file (optional)

Copy the URL below and share your Binder with others:

Fill in the fields to see a URL for sharing your Binder.

Copy the text below, then paste into your README to show a binder badge:
Basic  

*install.r*
*runtime.txt*
*apt.txt*

Slow but easy to setup. Recommended for beginners

[launch binder](#)
<table>
<thead>
<tr>
<th>Basic</th>
<th>free</th>
<th>Premium</th>
<th>free</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>install.r</code></td>
<td></td>
<td><code>Dockerfile</code></td>
<td></td>
</tr>
<tr>
<td><code>runtime.txt</code></td>
<td></td>
<td><code>install.r</code></td>
<td></td>
</tr>
<tr>
<td><code>apt.txt</code></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slow but easy to setup. Recommended for beginners</td>
<td></td>
<td>Faster launch</td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td>Premium</td>
<td>Pro</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td><strong>free</strong></td>
<td><strong>free</strong></td>
<td><strong>free</strong></td>
<td></td>
</tr>
<tr>
<td><code>install.r</code></td>
<td><code>Dockerfile</code></td>
<td><code>Dockerfile</code></td>
<td></td>
</tr>
<tr>
<td><code>runtime.txt</code></td>
<td><code>install.r</code></td>
<td><code>DESCRIPTION</code></td>
<td></td>
</tr>
<tr>
<td><code>apt.txt</code></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Basic**:
  - ![LaunchBinder](launch-binder.png)
  - Slow but easy to setup.
  - Recommended for beginners

- **Premium**:
  - ![LaunchBinder](launch-binder.png)
  - Faster launch

- **Pro**:
  - ![LaunchBinder](launch-binder.png)
  - Best for compendia
A fast set up binder

Dockerfile

DESCRIPTION

Pull a base image from Rocker (e.g., rocker:binder/latest)

rocker-project.org

The versioned stack

<table>
<thead>
<tr>
<th>image</th>
<th>description</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>r-ver</td>
<td>Specify R version in docker tag. Builds on debian:stable</td>
<td>239.7MB</td>
</tr>
<tr>
<td>rstudio</td>
<td>Adds rstudio</td>
<td>356.6MB</td>
</tr>
<tr>
<td>tidyverse</td>
<td>Adds tidyverse &amp; devtools</td>
<td>661.2MB</td>
</tr>
<tr>
<td>verse</td>
<td>Adds tex &amp; publishing-related packages</td>
<td>1GB</td>
</tr>
<tr>
<td>geospatial</td>
<td>Adds geospatial libraries</td>
<td>1.4GB</td>
</tr>
</tbody>
</table>
Include a workflow to manage relationships between data output and code.
drake
general purpose workflow manager & pipeline toolkit for reproducibility and high-performance computing.
github.com/ropensci/drake
Drake: *Data Frames in R for Make*

No cumbersome Makefiles

Vast arsenal of *parallel computing* options

Visualize dependency graph and estimate run times

Convenient *organization of output.*
Drake: visualize dependency graph
Take home

Leverage the R package structure and support tools/services as much as possible
Take home

Use modern tools to make your compendia more accessible, but don't forget long-term archives and simpler formats

github.com/topics/research-compendium
<table>
<thead>
<tr>
<th></th>
<th>Near term</th>
<th>Long term</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>piggyback, data packages</td>
<td>Zenodo and friends</td>
</tr>
<tr>
<td>environment</td>
<td>Binder and friends</td>
<td>Dockerfile</td>
</tr>
<tr>
<td>workflow</td>
<td>Drake</td>
<td>Core R tools, Make</td>
</tr>
</tbody>
</table>
KÖMPEINDIUM

1.

2.

3.

4.

5.