Fostering the next generation of sustainable software and reproducible open science

@_inundata
karthik.io
P values are just the tip of the iceberg

Peng & Leek, 28 Apr 15
50+ contributors • 
~65 software tools • R, C++, Node
Substantial contributor to CRAN
#1 Completing the data pipelines
Data retrieval (from APIs, data storage services, journals, and other remote servers).
Data visualization (interactive graphics in that extend beyond base tools).
Data deposition into research repositories, including metadata generation.
Data munging: With limited scope
Reproducibility (any tools that facilitate reproducible research, such as interfacing with git, tracking provenance or similar).
Open Science Tools
(Open science framework, GitHub-Figshare, Runmycode)

Open science Education
(Software Carpentry, NCEAS summer institute, D-Lab)

Domain Data repositories
(Dryad, GBIF, ITIS)

Domain Academics
(University researchers, federal scientists, independent researchers)

Informatics research
(NCEAS, BIDS, NESCent)

Industry
(Github, RStudio, Digital Science)
Text corpus

Public Library of Science (PLOS)
Biomed Central
eLife
Springer
IEEE
arXiv preprints
bioarxiv preprints
Elsevier (just kidding)

“The new science journalism and open science”
http://blog.revolutionanalytics.com
Data viz

```r
ee_observations(genus = "lynx") %>%
  ee_maps
```
DATA SHOULD BE machine readable
AUTOMATE BORING TASKS
Data publication

figshare
dvn
zenodo
dat *
+ EML
git2r

pipelines from * services

* Both in development
A quantifiable assessment of the capacity of our oceans to deliver benefits and resources sustainably.

Scores range from 0 to 100 for 10 goals for a healthy ocean.

Global Index Score

Ocean Region Rankings

New Score Releases
Antarctica and the Southern Ocean region scored 72 while the average score of the high seas was 67 out of 100.

Antarctica Scores
High Seas Overview
2014 Key Findings
#2
Building a user and developer community
Teach basic lab skills for scientific computing so that researchers can do more in less time and with less pain.

Teach basic concepts, skills and tools for working more effectively with data. Workshops are designed for people with little to no prior computational experience.
many transition from users to developers
Domain experts are often bad programmers
Only 19% of researcher contributed packages on have unit tests

46% of those rely on non-standard tests

Source: Oliver Keyes et al
rOpenSci Onboarding

Thank you for considering submitting your package to the rOpenSci suite. In order to ensure a consistent level of software quality for our users, we ask that you go through an onboarding process so we can better understand what your software does, and how we can help improve it before adding it to our suite. This process also allows us to ensure that your package meets our guidelines and provides opportunity for discussion where exceptions are requested.

How to submit your package to the rOpenSci suite

- First consult our policies and packaging style guide to ensure fit and consistency.
- Next, open a new issue.
- Copy the template from Step 2 in CONTRIBUTING.md into the issue body, and fill it out.
- Once you have completed the template, submit the issue. Someone from our team will take a look within 5 business days and respond with next steps. Once your package is approved, we will provide further instructions on transferring your repository to the rOpenSci organizational account.
@karthik and @gavinsimpson you're awesome! #148

SimonGoring opened this issue on Sep 26, 2014 · 4 comments

SimonGoring commented on Sep 26, 2014

Thanks so much for all your help @karthik and @gavinsimpson. This package probably could have limped on for a while, but it's really starting to look very polished, and I'm starting to feel much more confident in my own coding, even if some of these discussions make me feel like I'm barely treading water!

SimonGoring added the compliments label on Sep 26, 2014
Figure 4: Stratigraphic plot for Marion Lake. Age is plotted on the y-axis in calibrated radiocarbon years before present. The x-axis represents time in years. Cupressaceae, Tsuga, Alnus, Larix Abies, Other, Picea, Pinus, and Salix are plotted along the top axis.
Empowering the next generation of R developers
Hadley goes over data ingestion, including from flat files, SAS files, databases and web APIs

Moments [click on the time to jump to it]:
- 0:00 Hadley's talk
- 40:29 General question session starts
- 52:00 - Wrap up
Simon Goring
Simon is a postdoc at the University of Wisconsin, working on paleo-ecology. Simon maintains neotoma.

David Winter
David is a recent PhD grad from University of Otago, working on evolutionary genetics. David maintains rmentr.

Lee Marsh
Lee is a developer at the US National Phenology Network. Lee contributes to mpn.

Sara Varela
Sara is a biologist at the Department of Ecology of the Charles University in Prague. Sara maintains paleobioDB and Ravis with Javier.

Vijay Brave
Vijay is a graduate student at University of Kansas. Vijay contributes to rgbif and rvertnet.

Rafael Maia
Rafael is a graduate student at University of Akron, working on development and evolution of iridescent colors in feathers. Rafael maintains rebird.
“Absolutely stellar gender diversity” I think this is the most gender-balanced tech conference I’ve been to. Great diversity in career and academia too.”
Cultivating contributors
Cultivating contributions

Cumulative New Contributors

October 2010 to December 2014

2011/01  2012/01  2013/01  2014/01  2015/01
A RESEARCH PAPER

THE NARRATIVE

THE DATA

THE CODE
A RESEARCH PAPER

THE NARRATIVE

THE DATA

THE CODE
https://rOpenSci
@rOpenSci