PLOS Biology

rOpenSci - Open Tools for Open Science

--Manuscript Draft--

Manuscript Number:	
Full Title:	rOpenSci - Open Tools for Open Science
Article Type:	Presubmission Inquiry
Corresponding Author:	Karthik Ram, Ph.D. UC Berkeley Berkeley, CA UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	UC Berkeley
Corresponding Author's Secondary Institution:	
First Author:	Karthik Ram, Ph.D.
First Author Secondary Information:	
Order of Authors:	Karthik Ram, Ph.D.
	Scott Chamberlain
	Carl Boettiger
	Edmund Hart
Order of Authors Secondary Information:	
Abstract:	Reproducibility is a hallmark of scientific discovery, yet results from a majority of currently published work remain irreproducible [1]. A source of this problem is that the underlying data and code are often not shared alongside the paper, making it impossible to validate the results. Recent changes to journal guidelines and mandates from funders [2] are however addressing this problem at the policy level. Scientists still need tools that integrate easily into their existing workflow. This allows them to not only deposit data and code in public repositories, but also benefit from shared data to reproduce exisitng results and carry out novel synthesis research. To address part of this problem, we created a software collective called rOpenSci (http://ropensci.org/). R is open source software widely used in the academic community for data manipulation, statistical analysis, and visualization. However, R can also directly integrate with numerous databases and journals via application programming interfaces (API), making it possible to query, download and deposit data directly from R. Thus, scientists who synthesize novel results from existing data can share the entire workflow, from data acquisition to final results, as a single R script alongside the paper. Similarly, scientists generating original data can also submit those directly to a persistent repository after it has been analyzed in R. Our collective builds open source tools (referred to as packages in the R context) that facilitate this process. While the open science philisophy has gained considerable support in recent years, there is an equally important need for the right tools to facilitate the process. In addition to describing the tools we have developed thus far, we also describe our future directions and how they fit within the broader context of open science. [1] Costello, M.J. (2009). BioScience, 59, 418-427. [2]: "http://www.nsf.gov/pubs/2013/nsf13004/nsf13004.jsp?WT.mc_id=USNSF_109"

November 5th, 2012

Editor-in-chief
PLOS Biology

Dear editors,
My coauthors and I would like to propose an article for the **Community Pages** section titled **"rOpenSci - Open Tools for Open Science"**. In this article, we discuss a new approach to the challenges of reproducibility and data synthesis in the ecological sciences. Our approach emphasizes (1) building on tools that are already part of existing analytic workflows and (2) building a community of practice to facilitate adoption of more reproducible and scalable research practices.

The issue of reproducibility in science (or lack thereof) has received considerable attention in recent years. Several high profile papers have recently been retracted, and the results of a large majority of published articles remain irreproducible (Costello, 2009; Van Noorden, 2011). There are several reasons for this pervasive problem, some of which are technological (lack of tools or appropriate venues to deposit data and code) while others stem from the deeply rooted culture that rewards protecting data and code as trade secrets (Jones et al., 2006). Even in cases where scientists are in favor of data/software sharing, the technological barriers associated with using such tools remains high. However more journals are requiring reproducibility and data sharing, making accessible tools ever more important (Peng, 2009).

Although there are numerous tools that allow researchers to capture and share their workflow, many of these come with a significant learning curve and remain incompatible with the ones that scientists already use (Curcin et al., 2008). This presents a significant barrier to adoption. Among the tools used for statistical analysis and data visualization, R is one of the most widely used ones in academia, its popularity driven by the fact that it is open source and can easily be extended with user contributed packages (currently over 4000). Since the language is already familiar to a large portion of the scientific community, new tools built in R that facilitate reproducibility (by retrieving data associated with a publication and running the associated code), reuse, and discovery could become widely adopted with a very low barrier.

To address some of these challenges, we (all postdocs in ecology and evolution) began a software collective in 2011 called **rOpenSci**, short for R Open Science http://ropensci.org. We create tools that make it easy to discover and retrieve existing ecological and evolutionary data, mine full-text literature, both of which can facilitate reproducibility, novel discovery, and synthesis. Most importantly, these tools work in an environment that is already familiar to most scientists. Further, some of our tools also enable scientists to submit data to persistent repositories such as figshare and member nodes of DataONE (e.g. Data Dryad, OneShare). Scientists no longer need to pull data from different websites in disparate formats, which remain the key steps that are omitted in methods sections. Our tools integrate data discovery and access into a single common and widely used environment and the entire workflow can be shared easily because it is programmatic. These tools are very timely given NSF's recently announced changes to their merit guidelines where they now consider data and software as research products on par with publications.

We propose an article reviewing the importance and benefits associated with sharing data and code and the rationale and motivation behind our efforts. We will also briefly summarize the tools we have developed along with a discussion of future directions. Our work has been featured on numerous blogs, frequently discussed on Twitter (and other social media) and was also was a runner-up in the PLOS/Mendeley API challenge held last year.

We thank you for your consideration.

Sincerely,
Karthik Ram, Scott Chamberlain, Carl Boettiger, and Edmund Hart.

References

Costello, M. J. 2009. Motivating Online Publication of Data. BioScience 59:418–427.

Curcin, V., & Ghanem, M. 2008. Scientific workflow systems - can one size fit all?. 2008 Cairo International Biomedical Engineering Conference 1–9.

Jones, M. B., Schildhauer, M. P., Reichman, O. J., & Bowers, S. 2006. The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere. Annual Review of Ecology, Evolution, and Systematics 37:519–544.

Peng, R. D. 2009. Reproducible research and Biostatistics. Biostatistics (Oxford, England) 10:405–8.

Van Noorden, R. 2011. Science publishing: The trouble with retractions. Nature 478:26–8.

rOpenSci - Open Tools for Open Science

Abstract
Reproducibility is a hallmark of scientific discovery, yet results from a majority of currently published work remain irreproducible (Costello, 2009). A source of this problem is that the underlying data and code are often not shared alongside the paper, making it impossible to validate the results. Recent changes to journal guidelines and mandates from funders are however addressing this problem at the policy level (US NSF - Dear Colleague Letter - Issuance of a new NSF Proposal & Award Policies and Procedures Guide (NSF13004), n.d.). Scientists still need tools that integrate easily into their existing workflow. This allows them to not only deposit data and code in public repositories, but also benefit from shared data to reproduce exisitng results and carry out novel synthesis research.

To address part of this problem, we created a software collective called rOpenSci (http://ropensci.org/). R is open source software widely used in the academic community for data manipulation, statistical analysis, and visualization. However, R can also directly integrate with numerous databases and journals via application programming interfaces (API), making it possible to query, download and deposit data directly from R. Thus, scientists who synthesize novel results from existing data can share the entire workflow, from data acquisition to final results, as a single R script alongside the paper. Similarly, scientists generating original data can also submit those directly to a persistent repository after it has been analyzed in R. Our collective builds open source tools (referred to as packages in the R context) that facilitate this process. While the open science philisophy has gained considerable support in recent years, there is an equally important need for the right tools to facilitate the process. In addition to describing the tools we have developed thus far, we also describe our future directions and how they fit within the broader context of open science.

References

Costello, M. J. 2009. Motivating Online Publication of Data. BioScience 59:418–427.

US NSF - Dear Colleague Letter - Issuance of a new NSF Proposal & Award Policies and Procedures Guide (NSF13004). n.d. US NSF - Dear Colleague Letter - Issuance of a new NSF Proposal & Award Policies and Procedures Guide (NSF13004).