Project Summary

Science is in the midst of a dramatic transformation that is being driven by increasing access to large amounts of heterogeneous data. The long-established model where researchers collect and analyze their own data will soon be replaced by one where disparate datasets are brought to bear on both basic and applied problems. As science becomes more data-driven, it faces a whole new set of challenges. Researchers will not only have to maintain expertise in their domains but also learn new skills to curate, retrieve, and analyze these newly available data. Recent efforts such as DataONE, Data Dryad, and the Berkeley Initiative for Global Change Biology (BIGCB) have been developing the infrastructure needed for long-term storage and retrieval of environmental data, yet the tools and culture that help researchers become consumers of big open data are lacking.

To foster and support a new generation of data-driven science, we are proposing an integrated effort to build tools and training using our own community, Ecology and Evolution, as a model. Since mid-2011, we have been operating a software collective called rOpenSci to develop tools in R, the open source statistical environment that has already become the *lingua franca* of science. The suite of tools we have already built (more than 30 R packages) facilitate access to a wide variety of data repositories and lower barriers for use. We seek funds to support one full-time and two part-time data scientists and cover the cost of hackathons and travel to meetings for training and outreach in order to broaden the scope of our efforts and achieve four key objectives.

a) Focus on identifying shortcomings, strengthening our core products, and working to link existing tools through interoperable data structures and visualization routines. To facilitate this process, we will convene two hackathons, one with the core development team and advisors and a second with a wider group of rOpenSci collaborators.

b) Build a general purpose toolkit that will allow researchers in any community to develop their own tool to access domain specific data. This toolkit will help researchers access data sources available via application programming interfaces (API) and other, less-structured data.

c) Conduct outreach and training workshops at universities and conferences. In addition to increasing awareness of our tools and growing our user base, we will also use this opportunity to assess challenges and barriers to data science in closely related fields. As part of this process, we will interact with scientists outside of Ecology and Evolution to identify disciplines that could immediately benefit from more data science tools and training.

d) Explore long-term sustainability models for the project. This first year of funding will allow us to explore options, including strengthening the position of the organization to solicit similar funding from foundations and also pursue grants through National Science Foundation's Office of CyberInfrastructure and related agencies while continuing to operate within the confines of academic institutions. We will also explore establishing a non-profit, either independently or in collaboration with a university.

Data-driven science in Ecology and Evolutionary Biology (EEB)

The world of science is experiencing a data revolution. A majority of scientific discoveries of the 20th century came from datasets small enough to fit inside field and laboratory notebooks. In recent years, however, the size of these datasets has grown exponentially, a trend catalyzed by widespread availability of inexpensive sensors and storage technology. As a result, many new insights will come from vast stores of existing data, rather than from new data collection efforts. Jim Gray described this phase as the fourth paradigm in science (Bell et al. 2009; Tolle et al. 2011). Scientific training and practice have not kept pace with these developments, leaving this wealth of emerging data inaccessible to most researchers.

In order to fully realize the potential of data-driven science and allow researchers to draw insights from these vast data stores, we need to address challenges associated with all aspects of the research life cycle. Defined broadly, these activities include data capture, curation, and analysis (Tolle et al. 2011). The value of curating and preserving large databases has long been recognized since the creation of GenBANK, a genetic sequence database, in 1982. In the natural sciences, several recent projects, such as DataONE, Data Dryad, Berkeley Initiative in Global Change Biology, and NEON, are building the infrastructure necessary to curate and distribute data coming out of experiments, observations, and sensors. However, infrastructure development alone cannot drive the process forward. As evidenced by GenBANK's slow growth from 1983–1996, greater buy-in by researchers to share data and the tools used to analyze them is critical. Open and easy access to data and tools can be a research accelerator enabling scientists to rapidly collaborate on knowledge creation and synthesis efforts (Neylon 2012).

Technological advances, such as the digitization of museum specimens and field station records (Losos et al. 2013), are liberating vast quantities of data that were previously inaccessible and bringing together newer, more granular data from inexpensive sensors. The possibilities for new research questions that can rapidly advance science are endless when experimental data are combined with past records and environmental data measured at large spatial and temporal gradients (Wolkovich et al. 2012). Although many pieces necessary to bring the ecological sciences into the era of networked science are falling into place, other critical components are still missing (Hampton et al. 2012).

The incentives to share data are changing slowly in the research community. First, changes to merit guidelines by major funders and new ways to track metrics of productivity are driving data sharing. Changes to NSF merit guidelines now allow users to include publications and other products such as data and code that emerge from research endeavors ("US NSF - Dear Colleague Letter - Issuance of a new NSF Proposal & Award Policies and Procedures Guide (NSF13004)" 2012). Many journals also require authors to deposit data associated with publications either immediately or after a period of embargo (Whitlock et al. 2010; Fairbairn 2011). Finally, new NSF proposals also require an explicit data management plan (Donnelly et al. 2010). Together, these ensure that there will be more buy-in to depositing data into a service that will be machine readable.

Until recently, researchers who put time and effort into sharing well-curated data and well-documented code with unit tests were considered extreme altruists. Now, the scholarly landscape is in the midst of a revolution, and among the things that are changing are the incentive mechanisms. For example, altmetrics (Piwowar 2013), the use of social media to track influence of research outputs and data products, are driving more ways to measure success. Organizations and repositories including DataCite, Figshare, Dryad, DataONE, and others allow data that can be cited independent of publications. Papers that share data are more likely to receive citations (Piwowar et al. 2007), and people who collect and deposit well-curated data can receive measurable recognition for their efforts. This is especially important as the scientific community is calling for data citation to be part of the tenure and promotion practice (Parsons et al. 2010).

Data-driven science in ecology and evolutionary biology is still in its infancy, whereas industry and other scientific disciplines such as astronomy have been leveraging the power of big data for quite some time. We are slowly getting all the pieces together to build what Michel Nielson best described as the **well structured information commons** (Nielson 2011). However, the infrastructure developed thus far supports data archiving and distribution; now that data are online, we need to make the tools to consume them.

Why R?

R (http://r-project.org/) is an open-source statistical environment that can be used for statistics, data acquisition and manipulation, modeling, among other uses. R is increasingly being used by scientists across all disciplines and has overtaken popular scientific programming tools. Part of the reason behind R's explosive growth is the ease with which the ever-growing userbase can add new functionality, a fact evidenced by

5,000+ currently available R packages (Smith 2010). Many of these packages were developed by researchers to address domain-specific needs.

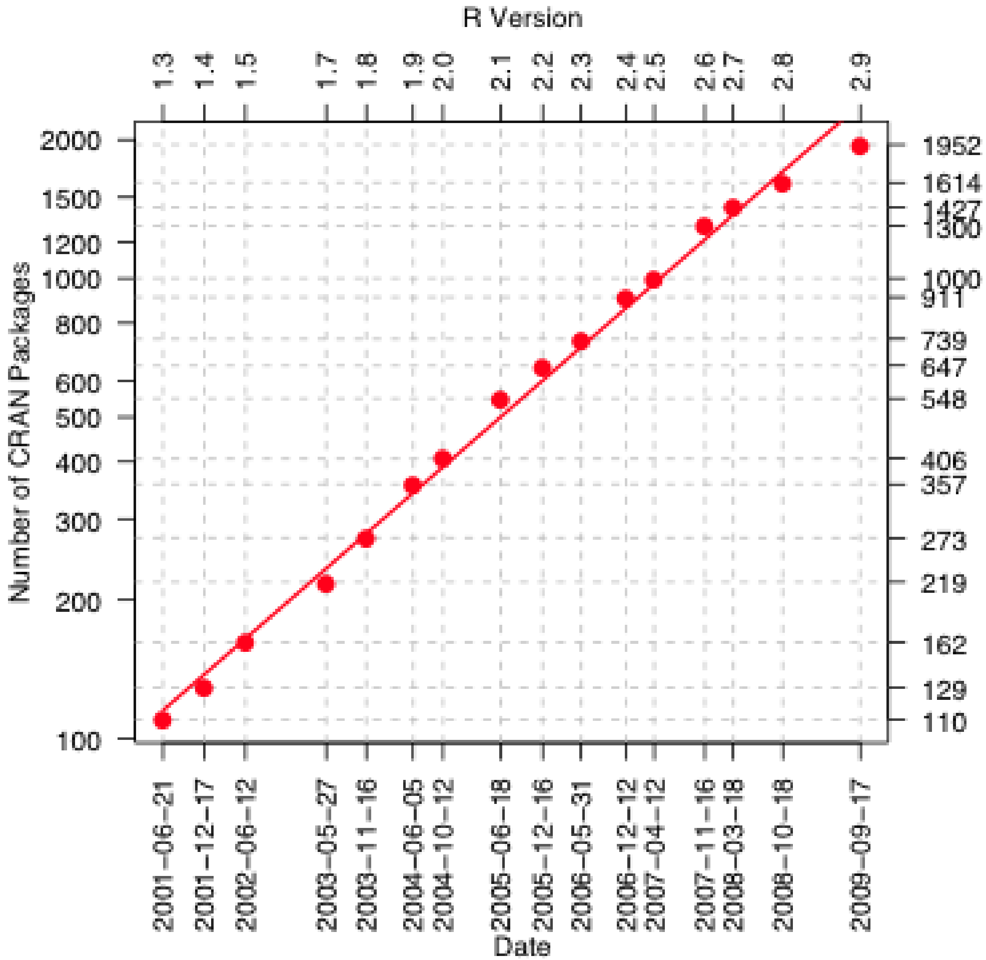

Figure 1: Exponential growth in availability of user contributed R packages with each new version of R. Source Smith, 2010

Although the popularity of R is difficult to measure with any single metric, several sources provide consistent evidence to support this claim (Vance 2009). The TIOBE community index, which tracks all programming languages, lists only two statistical tools (R and SAS) in the top 100. Discussions on message boards, blogs, and QA sites like Quora and StackOverflow can also provide strong proxies for interest and activity around any software tool (Schwartz 2009). All of these metrics place R well above competing tools, including Stata, SAS, and SPSS. Perhaps the strongest evidence to support R's role as powerful tool for data analysis comes from the fact that more than half of the competitors in data competitions such as those run by Kaggle use R (Muenchen 2012).

Because R, and its plethora of tools that can readily manipulate and analyze data, has a strong foothold in the scientific community, we believe that providing additional tools to consume data in the same environment would be a powerful catalyst to advance research. New R users can easily get support at their home institutions or on one of many active message boards. All of the packages we have developed carry a Creative Commons Zero `CC0` license (our Dryad package was the first R package on Comprehensive R Network CRAN to offer this license). Such open and permissive licenses are also key to accelerating open science (Neylon 2012).

Work to date

Since July 2011, we have developed a suite of tools that access ecological and evolutionary datasets and other data sources. We currently have more than `30` packages that fall under four broad categories:

- **Data packages**: Packages that interface with data repositories. These packages retrieve entire datasets such as data associated with publications, biodiversity occurrence data, or information needed to clean messy datasets (e.g. taxonomic resolution). Some specific data repositories include the Global Biodiversity Information Facility (GBIF) and Dryad.
- **Literature packages**: Packages that interface with full text of some journals and scholarly metadata. During the course of this grant, we will complete this suite with interfaces to other journals from publishers such as Springer, Frontiers, eLife, and PeerJ. Databases such as Mendeley, which allows for rapid literature discovery based on reference manager data from over two million users, are also included.
- **Altmetrics packages**: Packages that retrieve raw metrics from data providers (e.g., GitHub, PLOS) or aggregated metrics from major altmetrics providers (ImpactStory, Altmetric).
- **Hybrid packages**: To enable reproducible research, researchers also need to share their datasets and related products publicly. We provide access to two significant data repositories, **Figshare** and **DataONE member nodes**. The latter is still in early development, but we will play a key part in maintaining the package after the recent initial release.

We have two publications directly from the project (Boettiger & Temple Lang 2012; Boettiger et al. 2012) and will continue to publish vignettes in R and domain journals. Although we engage many users via social media, publishing on our R packages will ensure that we are able to reach researchers who are not active on social media.

We have developed a modern website http://ropensci.org/ that serves as the main point of engagement for our user base and includes a growing set of package tutorials.

We placed 3rd out of 40 entries in the Public Library of Science (PLoS) - Mendeley API binary battle, in which we showed off packages that interact with the PLoS and Mendeley APIs, both of which have gained traction with scientists.

We know that at least one of our packages is heavily used by the data provider, rplos. Our rplos package provides an R interface to the PLoS search and altmetrics APIs. Martin Fenner, the lead developer on the altmetrics API, uses rplos to consume data for internal PLoS analyses, evidence that our software is of great use to both academics and developers in the enterprise.

Objective 1: Technical development

We have already developed a suite of tools that access ecological and evolutionary datasets and others that span the natural sciences. A full list of packages that have been publicly released or in development is available at ropensci.org/packages/

We will continue development of our packages but also build a general purpose toolkit (see Objective 2) that will enable researchers to build their own R packages to interact with a data resource on the web.

We will focus on the following tasks in our technical development:

- **Strengthen our core suite**: We will continue to strengthen our core offerings. These include completing popular packages such as rDryad, rMendeley, and rPLoS. Currently there are several new API methods that we have not implemented due to lack of time. In addition to increasing the range of methods available through existing packages, we will also work on improving speed and efficiency by working with data providers to cache data when necessary. Several providers have expressed interest in working with us on this issue.

- **Standardize data structures**: We currently derive data and offer it to our users in the ad hoc formats determined by individual data providers. These include `XML`, `JSON`, `data.frame`, and `list`. We plan to spend time to make these formats interoperable, such that data downloaded from one package can be easily merged or joined with data from another package (where possible). This step would encourage our users to draw insights with data from multiple packages.

- **Develop visualization routines**: Visualizations are key to understanding data. Up to now, we have focused on bringing scientific data into R, and we will spend significant time developing easy to use visualizations for various data types, including maps, altmerics data, and literature metadata as generic R methods.

We plan to hold two hackathons, bringing together the core team and enthusiastic volunteer developers, that will serve two central purposes. The first is to improve existing package functionality and outreach within our current realm of scientific expertise. The second is to engage developers from other disciplines, allowing us an opportunity to interact with new potential collaborators, develop new package functionality, identify new data sources, and engage a broader community of scientific disciplines.

In the month preceding a hackathon, we will plan activities to ensure the best use of meeting time. These include brainstorming ideas and encouraging as many as possible without passing judgment. Once ideas are on the table, we will evaluate, rank, and prioritize them. At the meeting, we will break off into small groups and spend the morning designing and prototyping each one. After each group presents its efforts, we will move to implementation and testing. By day two, we will reevaluate each idea based on prototypes and design a plan to scale successful ideas across our entire suite. Other hackathon activities will include discussions of manuscripts and effectiveness of outreach campaigns.

Our efforts are key towards driving data science forward in the research community. The significant technical developments on the R side are coming from industry (Rstudio, Revolution Analytics), while scientists are building higher level wrappers. We play a unique role in this space in which we are able to transfer industry-standard tools to academic applications.

Since its beginning, rOpenSci has followed an agile development strategy. RStudio's `devtools` package makes it possible for us to rapidly deploy any package as soon as one or more functions becomes operational. With the large network of followers we have built up through social media and R-bloggers, we are able to get rapid feedback from early users and integrate continuously to work towards stable versions.

Objective 2: General purpose toolkit

While we plan to strengthen our core offerings, we will shift part of our focus away from developing more packages that address individual data sources to building a generalized toolkit that would allow our growing community of developers and data evangelists to contribute its own tools. This toolkit will (a) facilitate synthesis of data drawn from a diverse collection of repositories by providing a standardized framework, and (b) enable researchers to more easily develop their own software packages that integrate with online data repositories for which no existing R interface exists.

The toolkit will consist of:

- Tutorials for consuming web resources: These resources will allow anyone with knowledge of R to build a tool for a specific data source. The recent development of R packages `devtools` (for creating R packages), `testthat` (for testing R packages), and others make creating an R package relatively easy. However, interacting with web resources requires a different set of skills than typical R programming. We will develop a set of tutorials to teach users how to write R packages and functions to use web resources, including scraping data from websites, downloading compressed files, and using SOAP and REST APIs.
- R package templates: We will create a series of R package templates in repositories on GitHub that anyone can use to create R packages to interact with data repositories.

We will use this general purpose toolkit during our training sessions and workshops and use the feedback to bootstrap development throughout the year.

Objective 3: Outreach, community engagement, and community development.

Engaging the broader scientific community to encourage knowledge discovery through existing data sources using our tools is a central goal. Beyond just building tools, we want to effectively communicate with the scientific community about our tools, how they can use them, and how they can collaborate with us on tools for data sources we haven't yet developed.

We face two central challenges in outreach. The first is how can we further engage our existing users within the scientific community where we have the greatest traction: ecology and evolutionary biology (EEB)? The second is how can we identify and make inroads in other scientific disciplines outside of our roots, such as genetics, paleontology, conservation, etc.? We can address both of these challenges through the use of social media, workshops, and hackathons. These strategies will help us increase our profile within EEB, as well as identify key developers and collaborators outside of EEB that we can engage with to establish ourselves within those disciplines.

Goal 1: Engaging our current scientific community

We intend to use our existing connections in the broader EEB community to solicit workshop proposals at 15 universities. We will base our workshop model on the successful work of Software Carpentry (Wilson 2006) and host one large hackathon where we invite programmers and collaborators from across the country to help with the technical development of our packages.

The core mission of Software Carpentry is closely aligned with our efforts. Their boot camps provide researchers with the basic computing skills necessary to appreciate the value our tools, and we will add to these skills by training people to access domain specific API's and data sources. We plan to collaborate with Software Carpentry at boot camps where host institutions request R and serve as instructors, followed by additional time with local departments (EEB, Economics, Geology) to discuss our efforts and how we can reach their communities.

We will implement conference workshops as events that last 2-4 hours after the conclusion of a day's activities and work with conference organizers to seek an appropriate venue and cover the cost of refreshments. Ideally, we will use this time to jump-start discussions about the use of R in attendee communities and map out roadblocks and potential solutions. With adequate planning and moderation, we will learn as much as possible about each community.

Goal 2: Developing communities outside of ecology and evolution

We are deeply aware of the data sources, policies, and practices in EEB, but we are less familiar with those of other disciplines. We can approach this by developing collaborative relationships with partners across disciplines such as conservation, genetics, sociology, and economics, and we have already begun this process. One such collaboration involves geneticist David Winters on the package `rentrez`, which searches the Entrez genome database. By hosting interdisciplinary hackathons with collaborators outside of EEB and attending non-EEB meetings (e.g., The American Geophysical Union meeting), we can develop our presence in the larger scientific community. Hackathons with key developers in other communities will help us identify existing needs in those communities and establish collaborations. Providing workshops at other related, but non-EEB specific, conferences will help us broaden our audience and establish relationships with key potential collaborators. We will also apply to run a SeSync working group, which will allow us to collaborate with economists and sociologists who work on those issues at the intersection with ecology. All of this builds on our existing collaborations, such as with paleontology, and we will pursue others once our funding becomes clear.

Objective 4: Sustainability

Our fourth objective is to establish the long-term sustainability of our software, our goals, and the rOpenSci project itself. While software plays a critical role in all aspects of scientific research, attention has been focused primarily on the development of novel tools rather than in providing long-term stewardship, development, and support for existing infrastructure (Bastow & Leonelli 2010). A growing recognition of these challenges (Stewart et al. 2009) has spurred recent calls from major funders, such as NSF Advances in Biological Informatics, Data Infrastructure Building Blocks, and Software Infrastructure for Sustained innovation, to address this deficit. To best enable us to pursue these funding sources either directly or through partnerships with nascent organizations such as the Institute for Sustainable Earth and Environmental Software, we will (a) continue to raise the visibility of our organization through workshops (objective 2) and outreach (objective 3), (b) develop an advisory board for the project to help identify and pursue opportunities for long-term support, and (c) pursue opportunities for networking and advice from groups and organizations sharing these objectives for software sustainability.

We will also explore the alternative of incorporating rOpenSci as a non-profit institution. Becoming a non-profit will allow us to not only pursue foundation funding and competitive government grants, but to receive donations from individuals and corporations. Non-profits that teach coding skill as we will, Ladies Learning Code, Code.org, and CoderDojo, receive significant funding from companies like Microsoft and Intel, suggesting a possible direction.

The rOpenSci project is already committed to best practices for sustainable software, such as open source software with permissive licensing, version management of code on a widely used open and collaborative platform (GitHub), clear documentation based on literate programming ideals, and the use of automated test suites. Not only do these practices help ensure the longevity of our own software, but they provide an example to others as we work to address the challenges of sustainable software. We will ensure that all of our software products meet the guidelines suggested by the Software sustainability institute (http://www.software.ac.uk/online-sustainability-evaluation).

Team

rOpenSci began as an organic effort in mid 2011 through discussions and conversations developed among Carl Boettiger, Scott Chamberlain, and Karthik Ram. Shortly after, we turned the project into a software collective to aggregate all of our efforts under a single umbrella. We have developed more than 30 packages to date involving 16 collaborators spanning various natural science disciplines. Edmund Hart joined the core team in fall 2012.

The core development team is comprised of four EEB postdocs who are passionate open science advocates and strong R developers. Collectively we have submitted 12 packages to the Comprehensive R Archive Network

(CRAN), two peer-reviewed rOpenSci publications (Boettiger & Temple Lang 2012; Boettiger et al. 2012), and have built a strong network within the R, open science, and EEB communities including several other Sloan funded efforts such as PLOS ALM, ImpactStory, and Software Carpentry. Two of us (EH and KR) actively engage with software training through Software Carpentry and will be able to evangelize this effort during our training endeavors. We are also involved with other data initiatives such as DataONE (KR and SC) which puts us in a strong position to successfully execute our objectives.

We actively engage in various community hackathons including recent ones such as the PLoS Altmetrics conference and hackathon, Phylotastic a project to create tools for consuming phylogenetic trees, and EcoHackSF.

We have established an advisory board with leaders from ecological informatics (Matt Jones), computer science (Bertram Ludäscher) and the R core team (Duncan Temple-Lang and Hadley Wickham). More information about the board members is available at ropensci.org/about.

Assessment

To report measurable progress on our goals by the end of the year, we have developed the following tangible metrics that we could measure and report.

a. Acquire at least 30K unique visitors per year to the website as measured by our web analytics. We will host detailed tutorials on the website and track how often they are viewed or downloaded. Since we use GitHub to deploy packages and maintain master copies (in addition to the official R library repository - CRAN), we will submit a comprehensive impact report generated by ImpactStory. These metrics will also capture conversations about specific packages on social media channels. We will track GitHub statistics (watchers, stars, and forks) weekly to track temporal dynamics of package use.

b. Ensure that at least five papers (published or in press) cite one or more of rOpenSci's packages as a vital component of their methods sections. These five publications will not include any efforts that the core development team may publish. We will include letters from these authors evaluating the impact of our packages in their own research and how they perceive benefits to their colleagues and research communities. Our efforts are already being mentioned in recent publications (Poisot et al. 2013; Schäfer et al. 2013).

c. We plan to engage with researchers at five ecology and evolution meetings, two R and biostatistics meetings, and two interdisciplinary meetings. We will ask attendees to complete a before-and-after survey and follow up with them six months later to track shifts in their attitudes towards open science, open data, and programmatic access through R (Clark & Libarkin 2011).

d. Using our contacts and network in the R community, we will scope out potential future opportunities and identify ambassadors to lead such efforts. Through our interactions with R developers on StackOverflow, R-bloggers, and social media, we already have several such people who are leading efforts in literacy, paleontology, geology, and economics in mind.

Literature Cited

Bastow, R. & Leonelli, S. (2010). Sustainable digital infrastructure. *EMBO reports*, 11, 730–4.

Bell, G., Hey, T. & Szalay, A. (2009). Computer Science. Beyond the data deluge. *Science*, 323, 1297–8.

Boettiger, C. & Temple Lang, D. (2012). Treebase: an R package for discovery, access and manipulation of online phylogenies. *Methods in Ecology and Evolution*.

Boettiger, C., Temple Lang, D. & Wainwright, P.C. (2012). rfishbase: exploring, manipulating and visualizing FishBase data from R. *Journal of Fish Biology*.

Clark, S.K. & Libarkin, J.C. (2011). Designing a mixed-methods research instrument and scoring rubric to investigate individuals' conceptions of plate tectonics. *Geological Society of America Special Papers*, 474, 81–96.

Donnelly, M., Jones, S. & Pattenden-Fail, J.W. (2010). DMP online: the Digital Curation Centre's web-based tool for creating, maintaining and exporting data management plans.

Fairbairn, D.J. (2011). The advent of mandatory data archiving. *Evolution; international journal of organic evolution*, 65, 1–2.

Hampton, S.E., Tewksbury, J.J. & Strasser, C. a. (2012). Ecological data in the information age. *Frontiers in Ecology and the Environment*, 10.

Losos, J.B., Arnold, S.J., Bejerano, G., Brodie, E.D., Hibbett, D., Hoekstra, H.E., et al. (2013). Evolutionary biology for the 21st century. *PLoS biology*, 11, e1001466.

Muenchen, R. (2012). The Popularity of Data Analysis Software | r4stats.com on WordPress.com.

Neylon, C. (2012). Science publishing: Open access must enable open use. *Nature*, 492, 348–349.

Nielson, M. (2011). *Reinventing Discovery: The New Era of Networked Science*. Princeton University Press.

Parsons, M.A., Duerr, R. & Minster, J.B. (2010). Data citation and peer review. *Eos, Transactions American* . . . .

Piwowar, H. (2013). Altmetrics: Value all research products. *Nature*, 493, 159–159.

Piwowar, H.A., Day, R.S., Fridsma, D.B. & Fridsma, B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS One*, 2, e308.

Poisot, T., Péquin, B. & Gravel, D. (2013). High-Throughput Sequencing: A Roadmap Toward Community Ecology. *Ecology and Evolution*.

Schwartz, M. (2009). [SPAM Detected: 88%] Re: [R] R in the NY Times from Marc Schwartz on 2009-01-08 (R help archive).

Schäfer, R.B., Bundschuh, M., Focks, A. & von der Ohe, P.C. (2013). To the editor. *Environmental toxicology and chemistry / SETAC*, 32, 734–5.

Smith, D. (2010). The number of R packages is growing exponentially.

Stewart, C.A., Almes, G.T., Mccaulay, D.S. & Bradley, C. (2009). *Cyberinfrastructure Software Sustainability and Reusability Report from an NSF-funded workshop Report from an NSF-funded workshop held 27 and 28 March 2009*.

Tolle, K.M., Tansley, D.S.W. & Hey, A.J.G. (2011). The Fourth Paradigm: Data-Intensive Scientific Discovery.

US NSF - Dear Colleague Letter - Issuance of a new NSF Proposal & Award Policies and Procedures Guide (NSF13004). (2012). US NSF - Dear Colleague Letter - Issuance of a new NSF Proposal & Award Policies and Procedures Guide (NSF13004).

Vance, A. (2009). R You Ready for R? - NYTimes.com.

Whitlock, M.C., McPeek, M.A., Rausher, M.D., Rieseberg, L. & Moore, A.J. (2010). Data archiving. *The American Naturalist*, 175, 145–6.

Wilson, G.V. (2006). Where's the real bottleneck in scientific computing?. *American Scientist*.

Wolkovich, E.M., Regetz, J. & O'Connor, M.I. (2012). Advances in global change research require open science by individual researchers. *Global Change Biology*, 18, 2102–2110.