Open Science: research opportunities and incentive challenges

Carl Boettiger

Abstract

This whitepaper was written in response to seven questions on open science issued by invitation to the NSF/NIH ImagineU Conference (http://www.ncsa.illinois.edu/Conferences/ImagineU/) to be held on March 8-9, 2017. These answers reflect my own experiences and opinions at the time of writing, and focus primarily though not exclusively on examples from my own field of ecology and evolutionary biology, though I believe the trends and possibilities highlighted here are shared to great extent by other fields. This is a working paper whose primary goal is to spark discussion, not provide a comprehensive or authoritative viewpoint of the practice and potential of open science.

What is Open Science?

Open Science is the practice of striving for to make the process and products of research transparent and reproducible, including open data, open source software and code, open access publications, pre-prints, and open lab notebooks.

Simply put, open science is just science without the barriers created by other incentives. Galileo wrote of his discoveries in anagrams, such as: SMAISMRMILMEPOETALEUMIBUNENUGTTAUIRAS. This cypher allowed him to establish his priorty on the discovery of Saturn's rings without giving away its secret: he could later reveal the unscrambled solution, "Altissimum planetam tergeminum observavi" as evidence of his discovery. Soon his contemporaries would create the first scientific journals, taking the place of cryptic tricks to establish priority of discovery by openly sharing the details of one's study with colleagues world wide. The advantages of this more open system to the progress of science as a whole have been self-evident – academic journals have been the cornerstone of intellectual progress ever since. Yet Galileo's tension between discovery and recognition, between the needs of the research community and the needs of the individual – continue to exhert a potent and sophopriphic effect on the progress of science. Just as the advent of academic journals demonstrated in 1665, this inherit tension is not insurmountable provided the right combination of technological (publishing a copies letters from scientists in collections) and socio-economic change. The current movement of Open Science seeks to bring similarly transformative benefits to a process of science that today would otherwise appear terribly familiar to seventeeth century readers of *Philosphical Transactions*.

Open Science is an umbrella term that unites common threads of many areas, which for convience only I will divide into four main pillars: *Open Access*, *Open Data*, *Open Code*, and *Open Context*. Of these, *Open Access* is both the most mature and well-recognized movement since the advent of the internet in the 1990s, concerned primarily with removing paywalls charged by journal publishers. Solutions are usually divided into two alternatives: replacing reader (or more frequently, instituional) subscription charges with fees charged directly to the author (Article Processing Charges APCs, commonly seen in "Gold" open access), or in deposting pre-prints on a public archive (e.g. http://arXiv.org) prior to publication in a subscription journal ("Green" open access). More important but less recognized than the concern about *paywalls* is the concern about *licenses*: Open Access advocates recognize that the scientific literature is most useful when it is free from constraints on how text may be re-used in educational material, mined in large databases for common threads, and redistributed to its greatest impact. The Budapest definition http://www.budapestopenaccessinitiative.org/ closely aligns with the widley recognized Creative Commons Attribution (CC-BY) license, which grants explicit permission to re-use and and remix content. Much ink has already been spilt over the discussion of Open Access to the scientific literature, and the terms of the debate if not their resolution will be familiar to many. The predominant role played by established top-tier journals in hiring and advancement means that Open Access will remain a contentious issue for some time to come, until such journals decide they can viably convert their most coveted titles into an Open Access model

or changing metrics and expectations make them irrelivant. In this manner, it is precisely the same tension of personal recognition that scrambled Galileo's discoveries which continues hinder progress. Rather than rehash those issues here, I will focus primarily on other elements of Open Science which have recieved less attention but may have much greater potential for transformation.

Somewhere between Hooke's description of cells and the completion of the human genome, sceintific data outgrew what could be neatly written out in tables of a paper. The methods involved have also outgrown their medium: the description of an incline plane is one thing, the large hadron collider quite another. Computer analysis has come to play an ever more ubiquitous role in all aspects of scientific data collection, processing, and analysis, including the use of both existing software suites and custom code. As these elements of research have outgrown the printed page, they have become at risk to being lost to the scientific record all together. It is this separation from the academic paper that underlies the separate movements such as Open Data and Open Code. Separated from published articles, data can be distributed in more meaningful formats through central databases such as NCBI that can greatly faciliate preservation, discovery and distribution. Methods implemented in software can be applied reliably with only incremental effort.

Open Science is about more than access

Sharing isn't enough

Researchers face real but addressable technical barriers

In this piece, I summarize both the opportunities and challenges facing these elements of Open Science, before detailing my own attempts and compromises to navigate the tension between the public good and individual reality, and close with a few suggestions for easing the way forward.

The creation of scientific journals promoted three core principles which continue to underpin open science approaches today: *validation*, *scale,* and *novel insights.* By sharing descriptions of methodology as well as conclusions and results, the academic paper put reproducibility and validation of those results as the bedrock of scientific progress. By bundling the results of different scientists together in a periodical, journals brought a scale of information that no indivdual could rival through the previous mechanism of personal correspondance alone. And as Newton's most memorable quote reminds us, new insight has almost always been built upon reading, questioning and testing what has come before. These elements are so synonymous with the process of science that it easy to overlook how central a degree of openness is to these concepts.

Open Science: Open Access, Open Data, Open Code, and Open Context must still wrestle with the same tension of Galileo between promoting the needs of science: individual recognition and scientific progress.

they have also opened entirely new possiblities (see Reichman, Jones, and Schildhauer 2011) for advancing science by better leveraging that same collective enterprise first made possible by the creation of the scientific journal.

Promise and challenge of Validation

As that published literature grows ever faster, concerns about maintaining validation and scale, and yes, the novel insights that come with them, also mount. Replicating a study becomes ever more difficult and time consuming as data and methods, including software, have become largely external to the publication itself, (Garijo et al. 2013).

The promise and challenge of Scale

With the exception of certain special case applications of text mining (Van Noorden 2013), scientific literature is a fundamentally analog medium whose analysis cannot easily scale beyond the capacity of an individual's

eyeballs. In contrast, scaling is easy when data is accessible, discoverable, prepared and annotated in a machine-readable format. Sustainably engineered research codes and software with stable programming interfaces (APIs) and data formats can likewise interface with each other automatically without each researcher having to replicate methods already implemented by others.

Much as data scientists in industry and academia can leverage the digital debris left in in the wake of interent shopping and social media to train machines to translate languages and classify images, data collected in scientific research can be combined and re-used in ways never imagined. And data gathered from delibrate experiment and annotated and curated by experts can offer a signal-to-noise ratio that humbles terabytes from the click-stream. In every field of research, subdisciplines in informatics (such as in my own area of ecoinformatics, a la Jones et al. (2006)) have sprung up to adapt the advances from computer science and engineering to the problems of the discipline. This is giving rise to machine readable metadata descriptions such as the Ecological Metadata Langauge (Jones et al. 2006) and semantic descriptions such as OBOE (Madin et al. 2008) which can facilitate discovery and synthesis of related data from independent studies using highly heterogeneous data formats.

Like Open Access, Open Data is a movement in its own right, well established prior to widespread use of the umbrella term "Open Science."

Freed from the weight of historical precedent and established interests that have weighted so heavily on the Open Access debate, Open Data, Open Code, and Open Context offer a more unique possibility for rapid progress. The concern about paywalls, so central to the Open Access issue, causes little trouble in the context of Open Data, Open Code, or Open Context. Concerns about licensing are (rightly) more discussed but also closer to consensus. Without billion-dollar publishing houses holding the cards it is remarkable how easily the scientific community has converged on the issues of paywalls and licensing: most scientific data repositories provide content free of charge under permissive or public-domain licensing terms (though their own success could yet undermine that consensus). Yet on the other side of the same coin lies a major hurdle for these areas: as products of scholarship not typically considered in hiring and promotion, researchers may have little incentive to go through the effort and risk of preparing and sharing these outputs at all.

My open science practices

Must any open science practice go against the grain of individual interest?

My research, teaching and service emphasize open science practices. Yet look deeper and my decisions have still been shaped by tradeoffs and constraints.

Research

- **Publications**: All publications from my group are deposited in preprint servers and publisher's copies made available open access through the University of California Open Access Policy. Links to open access preprints are made available through the UC archive and my own website.

Few of my publications so far have appeared in (Gold) Open Access journals. Though my field has several reputable open access journals (and most journals offer a hybrid option of paying APCs), I have relied on preprint servers to provide open access copies of my papers. In some cases I have only archived preprint copies after a manuscript has been accepted, which serves the objective of Green Open Access (e.g. permissive/CC-BY licensing and paywall-free) though not the potential for additional feedback prior to publication. In my own experience I have found such unsolicited feedback rare; asking individuals directly for input is usually more successful.

In my experience I have not found alternative metrics to have a substantial impact on my career trajectory, though there is certaintly a limited extent to which I can assess what has mattered to those that have evaluated and hired me. While journal reputation seems to remain the leading currency of success, an interest in actual content still appears to be more valued than most metrics. Papers that have included Open Science

aspects such as providing accompanying software (Beaulieu et al. 2012; Boettiger, Coop, and Ralph 2012) or appealing to a larger audience (Boettiger 2015) have been better cited than the average paper in a top-tier journal (though my sample size should prevent us reading much into this).

- **Data**: Any original or processed data used in these publications is archived in an appropriate data archive, under a Public Domain declaration (CC0) with appropriate metadata and assigned a Digital Object Identifier (DOI). Small data files are managed directly in GitHub as `csv` text files with appropriate metadata in a README or Ecological Metadata Language (README file).

As a theorist I work primarily with "other people's data," which makes data sharing a relatively low bar for me – colleagues who have spent years fastidiously gathering their own data may naturally feel more possessive of it. Nevertheless, journal data archiving policies are starting to transform our field (A. J. Moore et al. 2010; Fairbairn 2011; Piwowar, Vision, and Whitlock 2011; Vines et al. 2013; Roche et al. 2015), to the point where today I believe our field does a much better job archiving data at time of publication than it does making use of archived data. Yet the great heterogeneity of data types, formats, and data management practices have limited how useful this data can be. Methods for effectively working with this very heterogenous data exist (e.g. EML, see Jones et al. (2006)) but outside of nationally funded efforts (LTER, NEON) with in-house informaticists these tools have not been widely adopted. Students and faculty frequently lack both training and tools to adopt best practices of data management (Hernandez et al. 2012; Cranston et al. 2014), let alone leverage emerging informatics techniques to benefit fully from access to machine-readable metadata when it is avaialble.

- **Code**: Code to reproduce the results of the analyses are made available through a public code repository under a permissive open source license (BSD-2) and a snapshot is archived in an appropriate academic repository with a DOI. Software developed in the course of an analysis is likewise archived and also actively supported & maintained (or deprecated when appropriate) through the appropriate software distribution archive such as Central R Archive Network (CRAN) for R packages, or Docker Hub for Docker images.

Code is a theorist's data. A nice simulation or algorithm could play a key role in many a paper to come, and I am not a stranger to the impulse to keep a particularly promising bit of code closer to the chest. On the other hand, authoring software has repeatably proven to bring it's own recognition should it truly prove useful to other researchers. I have written three explicitly software papers (Boettiger and Temple Lang 2012; Boettiger, Lang, and Wainwright 2012; Boettiger et al. 2015), and another two (Beaulieu et al. 2012; Boettiger, Coop, and Ralph 2012) owe many of their citations to software released along with the paper that has proven useful to other researchers. Software papers are largely a hack, an admission that we have allowed citation to assume a role of attribution over it's primary purpose of provenance. Nevertheless they are a very functional hack – many of the best-cited papers of all the past decades have been those describing software or algorithms, many of which are still cited long after the underlying software being acknowledged has changed. This is less than ideal for the purposes of provenance and reproduciblity – to that end, researchers should cite explicit versions of software directly. But on the other hand this serves as a very pratical mechanism to derive recognizable credit for the creation and maintenance of impactful software, whereas dividing citations over different versions or acrewing citations to an object many will still not consider in hiring and promotions does little good. While I am deeply supportive of software citation (https://www.force11.org/software-citation-principles), I believe the provenance of precisely what software a reseracher used will always be best communicated in sharing reproducible, scripted analyses directly, while software papers remain the most viable route we have to encourage open software development and maintenance.

A more subtle issue than simply sharing code or software under an open-source license is the consideration of what makes the software useful, trustworthy, reproducible, reliable, and sustainable in the long term. Best practices in documentation, unit testing & continuous integration, supporting bug reports and community contributions, presenting a standard and stable programming interface, choosing data structures & software dependencies that promote sustainability and platform compatibility are all elements that require additional training and additional effort to implement (G. Wilson et al. 2014). As the role of software written by academic researchers increases, our community is increasingly recognizing the importance of these practices (e.g. Joppa et al. 2013). With the right tools and training, these practices need not be an altruistic burden,

but can help the developer save time and effort down the road (Simperler and Wilson 2015).

- **Lab Notebook**: I have maintain an open lab notebook since 2010 at https://carlboettiger.info/lab-notebook. Since 2015 this has been maintained in a `notebook` directory of individual projects on GitHub, providing more granularity on projects & collaborations I have transitioned from student and post-doc to PI.

While my open lab notebook has been a visible element of my own open science practice (e.g. Wald 2010; Hayden 2013; Gewin 2013; Mascarelli 2014, Kitzes, J., Turek, D., & Deniz (2017)), I will not emphasize it here due to limitations of this approach for a broader community. While other aspects of Open Science focus on content made available with publication, open notebooks raise the issue of sharing content before publication. This raises the concern of scooping, already present in sharing data and code, to a higher level. In my own experience that concern is overstated, but a more important limitation of open notebooks is the issue of scale. Open Data and Open Software can scale better by leveraging well-defined standards.

Teaching & Service

I teach a graduate course, Reproducible and Collaborative Data Science and an undergraduate course: Data Science in Ecology and the Environment that both emphasize the practices, principles and tools of open science reproducible research. Additionally, I make all of my teaching materials for the course publicly available under an open, permissive CC-BY license with source code on GitHub repositories. My teaching methods and materials have also been greatly informed by the example and experience of other faculty that have openly shared course design and content, particularly Bryan (2016) and E. White and Brym (2016). I am senior fellow at the Berkeley Institute for Data Science where I am a member of the Open Science and Reproducible Research working group. The working group has just published a book on reproducible research in which I contributed a chapter on my current workflow (in Kitzes, J., Turek, D., & Deniz 2017). At a national level, I serve as a Science Adviser to the National Center for Ecological Analysis and Synthesis (NCEAS), which has been an early and important supporter of open science practices, including the Open Science for Synthesis program and an Open Science Codefest. I also serve on the User Board for the NSF Jetstream supercomputing center, where I try to serve as an advocate for an open and inclusive view of high performance computing towards domains not traditionally using HPC resources.

The way forward: Individuals must benefit from open science practices

The scientific journal replaced the anagram because it was in the best interest of the individual as well as the community at large to participate in this new approach. Open science practices advocated today cannot become widespread without a similar arrangement. This requires several types of change going forward. The first and most widely acknowledged is social change: the underlying incentive structure for research must shift. This issue dominates the attention of almost anyone seeking to reform academic research, but it should not be the only one. Changes at NSF such as the introduction of a Data Management Plan and the transition from listing "publications" to listing "products" on biosketches are both mechanisms and signs of this culture evolving, but cultural change on this scale will almost certaintly be a gradual process.

The other types of change are more immediately actionable. Both better technology and better training could help more individuals realize the benefits from open science practices rather than costs alone. The status emphasizes the altruistic benefits to the greater good over those to the individual in advocating for Open Access, Open Data, and Open Code. Sharing data or code at the time of publication is frequently seen as all cost and no gain: others might scoop your future results, you may be embarrassed by sloppy data managment or inelegant coding, and preparing data or code for distribution following best practices is both time consuming and unfamilar (Stodden et al. 2013).

Skepticism of technological solutions is well warranted (http://ivory.idyll.org/blog/2014-myths-of-computational-reproducibility.html) – many cyberinfrastructure projects that have promised technical solutions only to create signficantly underutilized and innaccessible platforms (Bucksch et al. 2016). Research software developers have

benefited significantly in recent years in advances from open source communities & platforms built by industry, such as GitHub, RStudio, and Travis-CI, and Docker; though tools developed primarly in academic context have also been very influential (including `knitr`, `numpy`, `juypter` & `matplotlib`). Still many of these tools are immature and further development is needed to adapt them to typical research workflow (https://ropensci.org/blog/2014/06/09/reproducibility/).

Reproducibility and validation must likewise be about opportunity rather than finger-wagging. Despite attention-grabbing headlines (Economist (2013)), issues of reproducibility and validation are not primarily a question of academic dishonesty but rather inherent to the complexity of scientific analysis. Silberzahn and Uhlmann (2015) provides a particularly illustrative study of this issue by comparing the results of 29 international research teams seeking to address the question: "are football (soccer) referees more likely to give red cards to players with dark skin than to players with light skin?" using the identical dataset. Small differents in statistical methods and in how the researh question is translated into testable model alter the conclusions reached by each team. Reproducible workflows are necessary if we are ever to evaluate the role such assumptions play, or to revisit earlier results as new data or methods become available.

Beaulieu, Jeremy M., Dwueng-Chwuan Jhwueng, Carl Boettiger, and Brian C. O'Meara. 2012. "Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution." *Evolution* 66 (8): 2369–83. doi:10.1111/j.1558-5646.2012.01619.x.

Boettiger, Carl. 2015. "An introduction to Docker for reproducible research, with examples from the R environment." *ACM SIGOPS Operating Systems Review* 49 (1): 71–79. doi:10.1145/2723872.2723882.

Boettiger, Carl, and Duncan Temple Lang. 2012. "Treebase: an R package for discovery, access and manipulation of online phylogenies." Edited by Luke Harmon. *Methods in Ecology and Evolution*, October, n/a–n/a. doi:10.1111/j.2041-210X.2012.00247.x.

Boettiger, Carl, Scott Chamberlain, Rutger A Vos, and Hilmar Lapp. 2015. "RNeXML: a package for reading and writing richly annotated phylogenetic, character, and trait data in R." *Methods in Ecology and Evolution*, n/a–n/a. doi:10.1111/2041-210X.12469.

Boettiger, Carl, Graham M Coop, and Peter Ralph. 2012. "Is your phylogeny informative? Measuring the power of comparative methods." *Evolution* 66 (7): 2240–51. doi:10.1111/j.1558-5646.2011.01574.x.

Boettiger, Carl, Duncan Temple Lang, and Peter C Wainwright. 2012. "rfishbase: exploring, manipulating and visualizing FishBase data from R." *Journal of Fish Biology* 81 (6): 2030–9. doi:10.1111/j.1095-8649.2012.03464.x.

Bryan, Jenny. 2016. "Data Wrangling, Exploration, and Analysis with R." http://stat545.com.

Bucksch, Alexander, Abhiram Das, Hannah Schneider, Nirav Merchant, and Joshua S. Weitz. 2016. "Overcoming the Law of the Hidden in Cyberinfrastructures." *Trends in Plant Science* xx. Elsevier Ltd: 1–7. doi:10.1016/j.tplants.2016.11.014.

Cranston, Karen, Luke J Harmon, Maureen A O'Leary, and Curtis Lisle. 2014. "Best practices for data sharing in phylogenetic research." *PLoS Currents* 6 (January): 1–8. doi:10.1371/currents.tol.bf01eff4a6b60ca4825c69293dc59645.

Economist. 2013. "Unreliable research: Trouble in the lab." *The Economist*, no. 19 October. http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.

Fairbairn, Daphne J. 2011. "The advent of mandatory data archiving." *Evolution; International Journal of Organic Evolution* 65 (1): 1–2. doi:10.1111/j.1558-5646.2010.01182.x.

Garijo, Daniel, Sarah Kinnings, Li Xie, Lei Xie, Yinliang Zhang, Philip E. Bourne, and Yolanda Gil. 2013. "Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome." Edited by Christos A. Ouzounis. *PLoS ONE* 8 (11): e80278. doi:10.1371/journal.pone.0080278.

Gewin, Virginia. 2013. "Turning point: Carl Boettiger." *Nature* 493 (7434): 711–11. doi:10.1038/nj7434-711a.

Hayden, Erika Check. 2013. "Mozilla plan seeks to debug scientific code." *Nature* 501 (7468): 472.

doi:10.1038/501472a.

Hernandez, Rebecca R, Matthew S Mayernik, Michelle L Murphy-mariscal, and Michael F Allen. 2012. "Advanced Technologies and Data Management Practices in Environmental Science: Lessons from Academia." *BioScience* 62 (12): 1067–76. doi:10.1525/bio.2012.62.12.8.

Jones, Matthew B, Mark P. Schildhauer, O.J. Reichman, and Shawn Bowers. 2006. "The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere." *Annual Review of Ecology, Evolution, and Systematics* 37 (1): 519–44. doi:10.1146/annurev.ecolsys.37.091305.110031.

Joppa, Lucas N., Greg McInerny, Richard Harper, Lara Salido, Kenji Takeda, Kenton O'Hara, David Gavaghan, and Stephen Emmott. 2013. "Troubling Trends in Scientific Software Use." *Science (New York, N.Y.)* 340 (6134): 814–15. doi:10.1126/science.1231535.

Kitzes, J., Turek, D., & Deniz, F., ed. 2017. *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences.* Oakland, CA: University of California Press.

Madin, Joshua S, Shawn Bowers, Mark P Schildhauer, and Matthew B Jones. 2008. "Advancing ecological research with ontologies." *Trends in Ecology & Evolution* 23 (3): 159–68. doi:10.1016/j.tree.2007.11.007.

Mascarelli, Amanda. 2014. "Research tools: Jump off the page." *Nature* 507 (7493): 523–25. doi:10.1038/nj7493-523a.

Moore, Allen J, Mark a McPeek, Mark D Rausher, Loren Rieseberg, and Michael C Whitlock. 2010. "The need for archiving data in evolutionary biology." *Journal of Evolutionary Biology* 23 (4): 659–60. doi:10.1111/j.1420-9101.2010.01937.x.

Piwowar, Heather A., Todd J. Vision, and Michael C Whitlock. 2011. "Data archiving is a good investment." *Nature* 473 (7347): 285–85. doi:10.1038/473285a.

Reichman, O.J., Matthew B Jones, and M. P. Schildhauer. 2011. "Challenges and Opportunities of Open Data in Ecology." *Science (New York, N.Y.)* 331 (6018): 692–93. doi:10.1126/science.1197962.

Roche, Dominique G., Loeske E. B. Kruuk, Robert Lanfear, and Sandra A. Binning. 2015. "Public Data Archiving in Ecology and Evolution: How Well Are We Doing?" *PLOS Biology* 13 (11): e1002295. doi:10.1371/journal.pbio.1002295.

Silberzahn, Raphael, and Eric L. Uhlmann. 2015. "Crowdsourced research: Many hands make tight work." *Nature* 526 (7572): 189–91. doi:10.1038/526189a.

Simperler, Alexandra, and Greg Wilson. 2015. "Software Carpentry get more done in less time," June. http://arxiv.org/abs/1506.02575.

Stodden, Victoria C, D H Bailey, J Borwein, R J Leveque, W Rider, and W Stein. 2013. "Setting the Default to Reproducible." *SIAM News*, 1–19.

Van Noorden, Richard. 2013. "Text-mining spat heats up." *Nature* 495 (7441): 295–95. doi:10.1038/495295a.

Vines, Timothy H, Rose L Andrew, Dan G Bock, Michelle T Franklin, Kimberly J Gilbert, Nolan C Kane, Jean-sébastien Moore, et al. 2013. "Mandated data archiving greatly improves access to research data." doi:10.5061/dryad.6bs31.

Wald, Chelsea. 2010. "Scientists Embrace Openness." *Science*, April. doi:10.1126/science.caredit.a1000036.

White, Ethan, and Zachary Brym. 2016. "Data Carpentry for Biologists: Teaching the Tools to Get Computers to Do Cool Science." http://www.datacarpentry.org/semester-biology/.

Wilson, Greg, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, et al. 2014. "Best Practices for Scientific Computing." Edited by Jonathan A. Eisen. *PLoS Biology* 12 (1): e1001745. doi:10.1371/journal.pbio.1001745.