Some thoughts on Open Science

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.

So what exactly does "open" mean?

I use the term "open" to mean something that is roughly consistent with the Budapest definition of "open" access, permitting readers to reuse and remix content, but will avoid any strict definition of what it means to be open, which is not only something of a continuum but also a multi-dimensional concept. This is most familiar and precise in the context of copyright licensing, e.g. CC-BY is consistent with this definition, though more restrictive CC licenses often used by other publishers falls short of it (e.g. by restricting reuse or remixing/re-purposing that is an important element of how science builds upon existing work). Software licensing uses a different vocabulary in regard to openness. Permissive licenses such as MIT and BSD are closest to the Budapest definition of Open Access, and are widely recommended by Open Science advocates (see Stodden 2009). For most purposes, the more general classification of any "Open Source" license, e.g., as defined by the Open Source Initiative https://opensource.org, widely recognized as the authoritative definition, also recognizes so-called viral or copyleft licenses such as the GPL which include a clause similar to the share-alike provision of Creative Commons. Openness in software raises issues that go beyond the domain of Copyright law, in particular, regarding the issue of patents. While the issue of viral/share-alike clauses is likely the most discussed divide among open source licenses, differences with respect to patent implications are at least as relevant: for instance, many universities, including the University of California system permit the use of both certain permissive and copy-left open source licenses (e.g. BSD-2, GPL-2) but forbid researchers to use others (e.g. the permissive Apache license or the copyleft GPL-3) ostensibly due to the potential patent rights that may be forfeited by the latter. Similarly, the lawyers and scholars of the Open Source Initiative declined approval of the Creative Commons Zero Public Domain Declaration, though it is recognized as a Free, GPL-compatible license by the Free Software Foundation, which oversees the GPL license over concerns about a clause which states it has no patent implications (see https://lists.opensource.org/pipermail/license-review/2012-February/thread.html). Elsewhere, journals (e.g. PLOS CompBio) and funders (e.g. National Science Foundation) recognize only licenses already recognized by OSI. **These examples illustrate that overly permissive terms may be just as much a barrier to reuse as overly restrictive licenses, and more generally, that precise and practical definitions of "open" are difficult**. In light of this, my definition of "Open" and "Open Science" should be seen as aspirational rather than prescriptive; a shared intent rather than shared practice. Here I share my vision of this shared intent, how it manifests in my research, and what are some of the major challenges and opportunities for open science that lie ahead.

Challenges and Opportunities

Open science advances and accelerates research and training by making it easier to verify, correct, and extend existing research and training efforts. Barriers to access any component of scientific research: conclusions, methods, data, software tools, etc, inherently slow the progress of science.

The best way to understand the importance of Open Science is to understand how these barriers currently impact research. This will help establish some broad context as to the present state of open science practices. While there is great heterogeneity in the manner and degree of open science practiced in different fields, countries, journals, and institutions, I believe that several broad generalizations are possible. Recognizing these generalizations first will permit me to focus my replies to these queries on those areas where the benefits have been least-well enunciated, accepted, or put into practice by our scientific communities.

I attempt to address these related questions on the *importance* and the *challenges* together:

1. Why is open science important for transforming research and learning?
2. What are the major technical, organizational, social, or cultural challenges you face, particularly as related to openness and sharing within your university and academia?
3. How can open science increase the societal impact of university research?

Despite the obvious synergies between the different elements of open science in scientific research as a process, I think it is essential to take the different components: open access, open data, open code, and open context, separately in order to understand the current challenges and opportunities each entails. The different components of open science are primarily important to somewhat different audiences, and have achieved different degrees of recognition. I address the challenges and opportunities of each pillar in turn here, deferring to other literature where the case or at least the terms of debate have already been clearly established in the community. The importance of at least two of these pillars of open science: open access and open data – have been so well established as to be now be mandated in some form by most major journals and funders (e.g Gewin 2016; Roche et al. 2015). The other two remain much more of the frontier.

Open Access

The terms of debate on open access and data archiving have been clearly articulated for some time throughout the scientific literature: after all, the importance of that literature as representing the progress and deliverables of science has never been in doubt. The primary barrier to open access today is the historical prestige of leading subscription-based journals, which researchers are reluctant to abandon and publishers reluctant to tamper with. Though new journals at all levels tend towards an open access model, the tight coupling of credit and prestige with established journals makes a rapid transition towards open access unlikely. Meanwhile, mandates by funders such as NIH, Burroughs Welcomme Fund, as well as institutional requirements such as the University of California Open Access Policy have instead largely promoted a model which splinters content into 'publisher' and 'open' versions through the use of preprint or institutional archives. Because institutions responsible for the lion's share of research dollars usually have subscription access to the most relevant journals, those with the most influence have the least incentive to change this system. Fortunately, some of the more transformative potential of open science lies in other venues that face barriers not quite so entrenched.

Open Data

While most journals have long required any supporting data to be 'available on request', many are becoming sensible to the well-documented (e.g. Roche et al. 2015 and references therein) practical limitations of this model and now require data archiving in public repositories. In some ways the situation here is better than that of journal articles: these data may be accessed without a paywall common to journal articles, and the cost to authors depositing data zero or at very least much less than typically found in article processing charges (APCs) of open access journals (e.g. http://dataDryad.org). No doubt this reflects the relative

absence of already-entrenched interests in academic publishing. Access to large pools of *scientific data* free of paywalls and copyright could be far more transformative than open access to scientific papers alone, as in the age of data science the processing of such data can *scale* through automation and algorithm in a way that its qualitative textual descriptions in papers never will. (e.g. see Reichman, Jones, and Schildhauer 2011). Further, scientific data accessible free from paywalls and copyright should be far easier to recombine in novel ways to pursue questions relevant to both academic and industry research.

However, technical and socio-cultural expectations have made compliance with these journal policies difficult and unpopular to enforce (or even measure), and have further limited the usefulness of the data even when it is deposited. Standards for organizing, annotating, and distributing scientific data are often immature or ignored by researchers never trained in the curation of necessary metadata, context, and formats that facilitating data preservation and reuse. Once again, academic incentives for success (almost exclusively measured in the number of high impact publications) are aligned against these practices, since they can take time away from other pursuits, may facilitate research competitors, and which bring little reward by themselves. Further, a lack of user-appropriate technology makes adopting best practices for data sharing extremely difficult. Likewise, a lack of technological tools that can make effective use of this highly heterogeneous data dramatically limits the scientific insights that could otherwise be gleaned from it. Despite a general consensus on when and how such published data should be cited, lack of training, precedent, and limits on bibliographies impede this practice.

As a result of these forces, we currently enjoy rapidly growing data archives but have been able to realize little of the transformative potential that truly open data can enable. In time, ever growing-archives, new learning and inference methods and the existence of some well-annotated data sets may overcome the lack of standards and metadata, but real progress will require greater recognition of data as a product of scientific research in order to promote higher standards of its preparation and curation.

Open Code

The remaining pillars of open science: (3) open code and (4) open lab notebooks and similar elements of the progress rather than products of science (reagents, protocols, and so forth) are significantly less discussed and developed into practices and policies of researchers, journals, funders and other institutions. Advances in technology and the rapid increase in scale of data available to many scientific disciplines has made *using* software an increasingly essential element of almost any scientific research, while also stimulating growing demand for *developing* software to address specific data and research questions. Open code is starting to receive the attention that Open Data has benefited from in the previous decade, with a handful of journals discussing or developing tentative policies regarding the availability and sharing of code (e.g. new policy at *Science* (Hanson, Sugden, and Alberts 2011), the rise of so-called 'software journals' such as PLOS Comp Bio, JOSS, JORS, and others).

Open code may be subdivided into "software," which I will use to refer to any code written with the intent that other researchers may use it as a tool to perform other research, vs "code" which researchers write in pursuit of their own research. This will always be more of a continuum than a binary division – in practice "code" that follows certain best-practices (functionalized, appropriate abstractions) can be very useful for further research, while some software development teams may have few real users.
(This division between intent is also present in data archiving, between large team efforts such as genome sequencing or NEON that seek to generate well-curated data for others, vs the long-tail of data being generated en route to an individual researcher's own objectives). The challenges and opportunities are somewhat different between these categories. With respect to open code, the primary objective of open science movement has been to encourage researchers to "publish your code, it is good enough" (Barnes 2010). Access to such code is often essential to replicating and understanding results drawn from it. In contrast, the emphasis on open software is that it has largely not been good enough – more attention to best practices of software development and open source community-building can improve the reliability, usability, provenance, integrity, and reproducibility of the software over the long term (e.g. see papers from the WSSSPE workshops, http://wssspe.researchcomputing.org.uk/).

Open context (open notebooks, open protocols, open reviews, etc)

These elements have received the least attention – indeed it is not clear that they should be should truly be lumped together. Yet each of these components face a more fundamental division from those previously discussed in that they shift the focus from something that can become recognized as a *product* in it's own right to something that is part of the *process*. Ultimately a shift from *product to process* captures the ideal of open science – by it's nature scientific research is always a work in progress which builds on previous products. The ability to share process directly both improves our ability validate and uncover limitations of previous work as well as extend and synthesize isolated results into general patterns.

How is open science part of, and important for your own research, teaching, and service agendas?

My research, teaching and service emphasize open science practices.

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.
- **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).
- **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.
- **Lab Notebook**: I 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.

My motivation and efforts to conduct my own research in an open science have also been discussed elsewhere (e.g. Wald 2010; Hayden 2013; Gewin 2013; Mascarelli 2014, Kitzes, J., Turek, D., & Deniz (2017)).

Teaching

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 White and Brym (2016).

Service

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.

If you had a senior leadership role in a university, what would you do to promote change and improve your university?

Open science practices often fall outside the standard academic scope of incentives and the currency of hiring and promoting. It is possible and can be effective to align these objectives whenever possible: for instance, software papers can translate software development into a more traditional currency of publication and citation, and certain grant opportunities may look favorably on open science goals and practices, and a researcher recognized for these practices may enjoy less tangible benefits of recognition. Yet for the most part these efforts will be at odds with the expectation to publish regularly in prestigious venues which makes up the cornerstone of scientific reward. The best thing senior university leadership can do to promote open science would be to protect and encourage these practices among junior faculty.

What risky and potentially transformative, big idea research proposal would you be writing if you had the right open science resources, and institutional support?

I firmly believe the most transformative big research idea would be the creation of the right open science resources, including institutional support, in the first place. While I will always be excited and fascinated by research in my own field in theoretical ecology and conservation decision-making and argue as vigorously as the next scientist about the fundamental importance of understanding the natural ecosystems that support our food security, economic and cultural stability, research which transforms how much larger parts of our scientific community generates knowledge and sees that knowledge translated into invention, policy, and application will by its very definition have greater and broader impact. Many of the most transformative changes to happen both in science and beyond in the past several decades have been the result not of demonstrating that something was possible for a small group of highly specialized individuals, but in demonstrating that it was practical at scales which made the previously inconceivable routine: high throughput sequencing, Moore's law, big data and machine learning. That the most transformative work increases the productivity of the greatest number of researchers (both in academia and beyond) is nearly tautological. Status quo practices that stand in contrast to open science practices are some of the most obvious and artificial sources of inefficiency in research productivity, since they only increase friction (or more precisely, prevent the decrease in friction due to technological advances) in access and exchange of scientific knowledge. Few if any of these practices prevent exchange and thus provide some potentially valuable protection (e.g. in the manner of a patent), but instead add friction in the form of paywalls, data that can only be accessed on request and manually converted to a machine readable form. Research data is most useful when it is preserved in open, machine-readable, well annotated formats. Results are most useful when they can reach the largest audience, be more easily validated, reproduced and extended. Research software will make little impact if it is poorly maintained or cannot be extended to changing types and sizes of data or integrated with other methods.

Only through education training of the next generation can we accomplish wide-reaching reform of such practices. We need a generation of scientists who both the knowledge and the desire to perform research in a manner that maximizes its potential impact. Basic training in broad-sense data science and reproducible research is fundamental for this transformation, both in academia and beyond.

Barnes, Nicholas. 2010. "Publish your computer code: it is good enough." *Nature* 467 (7317): 753.

doi:10.1038/467753a.

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

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

———. 2016. “Data sharing: An open mind on open data.” *Nature* 529 (7584): 117–19. doi:10.1038/nj7584-117a.

Hanson, Brooks, Andrew Sugden, and Bruce Alberts. 2011. “Making data maximally available.” *Science (New York, N.Y.)* 331 (6018): 649. doi:10.1126/science.1203354.

Hayden, Erika Check. 2013. “Mozilla plan seeks to debug scientific code.” *Nature* 501 (7468): 472. doi:10.1038/501472a.

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.

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

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.

Stodden, Victoria. 2009. “ENABLING REPRODUCIBLE RESEARCH : OPEN LICENSING.” *International Journal of Communications Law and Policy.*

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/.