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 the subject of his study: he could later reveal the unscrambled solution, "Altissimum planetam tergeminum observavi"", or"I have observed the highest planet tri-form" 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. <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.

Other elements of Open Science: Open Data, Open Code, and Open Context – still wrestle with same fundamental tension of Galileo between individual recognition and scientific progress. Yet freed from the weight of historical precedent and established interests that have weighted so heavily on the Open Access debate, these areas 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. Before I can discuss the opportunities and challenges presented by Open Data, Open Code, and Open Context of research, we will need a bit of background.

Somewhere in 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. Methods have also outgrown their medium: the description of an incline plane is one thing, the large hadron collider quite another. In particular, 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 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. These benefits can be loosely divided into three categories: *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. Yet as that published literature grows ever faster, concerns about maintaining validation and scale, and yes, the novel insights that come with them, also mount. As data and methods, including software, have become largely external to the publication itself, replicating a study becomes ever more difficult and time consuming (Garijo et al. 2013). 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.

Despite attention-grabbing headlines ((**???**)), 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.

The Promise of Open Data

Much as data scientists in academia & industry 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" (or "Open Research", which also embraces research in the humanities).

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.

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.

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.

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.

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

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