The Tao of Open Science for Ecology

Stephanie E. Hampton, Center for Environmental Research, Education and Outreach, Washington State University, Pullman WA 99164 s.hampton@wsu.edu

Sean S. Anderson, Environmental Science and Resource Management Program & Pacific Institute for Restoration Ecology, California State University Channel Islands, Camarillo, CA 93012 sean.anderson@csuci.edu

Sarah C. Bagby, Marine Science Institute and Department of Earth Science, University of California, Santa Barbara, 93106 bagby@geol.ucsb.edu

Corinna Gries, Center for Limnology, University of Wisconsin, Madison, 53706 cgries@wisc.edu

Xueying Han, Center for Nanotechnology in Society, University of California, Santa Barbara, CA 93106 xueyinghan85@gmail.com

Edmund M. Hart, National Ecological Observatory Network, 1685 38th St., Suite 100, Boulder, CO 80301 edmund.m.hart@gmail.com

Matthew B. Jones, National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, 735 State St. Suite 300, Santa Barbara, CA 93101 jones@nceas.ucsb.edu

W. Christopher Lenhardt, Renaissance Computing Institute (RENCI), University of North Carolina at Chapel Hill clenhardt@renci.org

Andrew MacDonald, Department of Zoology, University of British Columbia, Canada macdonald@zoology.ubc.ca

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

William K. Michener, College of University Libraries and Learning Science, MSC05 3020, University of New Mexico, Albuquerque, NM 87131 william.michener@gmail.com

Joe Mudge, The Institute of Environmental and Human Health, Department of Environmental Toxicology, Texas Tech University, Lubbock, TX 79416 joe.mudge@ttu.edu

Afshin Pourmokhtarian, Department of Earth & Environment, Rm 130, 685 Commonwealth Avenue, Boston University, Boston, MA 02215, apourmok@bu.edu

Mark Schildhauer, National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara 735 State St. Suite 300, Santa Barbara, CA 93101 schild@nceas.ucsb.edu

Kara H. Woo, National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, 735 State St. Suite 300, Santa Barbara, CA 93101 woo@nceas.ucsb.edu

Naupaka Zimmerman, School of Plant Sciences, University of Arizona, 1145 E 4th St., Tucson, AZ 85721, naupaka@gmail.com

Corresponding author: Stephanie Hampton, s.hampton@wsu.edu

PeerJ PrePrints

Abstract

The field of ecology is poised to take advantage of emerging technologies that facilitate the gathering, analyzing, and sharing of data, methods, and results. The concept of transparency at all stages of the research process, coupled with free and open access to data, code, and papers, constitutes "open science." Despite the many benefits of an open approach to science, a number of barriers to entry exist that may prevent researchers from embracing openness in their own work. Here we describe several key shifts in mindset that underpin the transition to more open science. These shifts in mindset include thinking about data stewardship rather than data ownership, embracing transparency throughout the data life-cycle and project duration, and accepting critique in public. Though foreign and perhaps frightening at first, these changes in thinking stand to benefit the field of ecology by fostering collegiality and broadening access to data and findings. We present an overview of tools and best practices that can enable these shifts in mindset at each stage of the research process, including tools to support data management planning and reproducible analyses, strategies for soliciting constructive feedback throughout the research process, and methods of broadening access to final research products.

Keywords: data management; ecology; open access; open science; reproducible research

Introduction

Ecology stands at the threshold of a potentially profound change. Ever-increasing computational power, coupled with advances in Internet technologies and tools, are together catalyzing new ways of pursuing ecological investigations. These emerging approaches facilitate greater communication, cooperation, collaboration, and sharing, not only of results, but also of data, analytical and modeling code, and potentially even fully documented workflows of the – processes—warts and all—that lead to scientific insights. This vision of free and unfettered access to all stages of the scientific endeavor has been called "open science" (Nielsen 2011). As an integrative and highly multidisciplinary field, ecology particularly stands to benefit from this open science revolution, and many ecologists have expressed interest in enhancing the openness

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

of ecology. To date, such conversations among ecologists have largely occurred online (e.g., discussed in Darling *et al*. 2013); thus it seems timely to present an introduction of open science for ecologists who may or may not currently be active in the social media forums where the discussion is evolving. We give an overview of the rise of open science, the changes in mindset that open science requires, and the digital tools that can enable ecologists to put the open science mindset into practice.

The exchange of scientific information was institutionalized in the 1660s with the establishment of the Philosophical Transactions of the Royal Society of London and the Journal des Sçavans, the first scientific journals (Beaver and Rosen 1978). While these journals provided platforms for scientists to share their results and ideas, they were largely accessible only to elites - those who could afford a subscription themselves, or those who belonged to an institution that held copies (Nielsen 2011). Individual scientists published in these journals to establish precedence of discovery; the notion of collaboration among scientists does not seem to have taken hold until the 1800s (Beaver and Rosen 1978).

The scientific world looks very different now. Advances in computing power and speed have accelerated not only individual scientists' discoveries but also their collaborative potential (Box 1). Modern scientists constitute a global college, its philosophical transactions enabled by the Internet (Wagner 2008), and collaboration has become the predominant norm for high-impact research (Wuchty *et al*. 2007). Technological developments also have enabled the capture (at ever increasing rates) of a previously unimaginable volume of data and metadata (Reichman *et al.* 2011, Dietze et al. 2013), and have underlain the use of increasingly complex models and analysis techniques to understand these data. Traditional paper notebooks cannot meet the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

PeerJ PrePrints

challenges of these new rates of accumulation, sharing, and recombination of ideas, research logs, data, and analyses (Strasser and Hampton 2012, Ince et al. 2012). The tools and approaches that together constitute open science can help ecologists to meet these challenges, by amplifying opportunities for collaboration and rewarding the creation of the consistent and machine-readable documentation that is necessary for reproducibility of complex projects.

While interest in this new paradigm is on the rise (Fig. 1), it must be acknowledged that both technical and sociocultural obstacles impede adoption for some ecologists. For example, precedence, attribution, investment, and payoff are high-stakes issues for professional scientists (Hackett 2005). Adopting open practices means ceding some control of these issues, learning new standards and practices for exerting control over others, and devoting precious time to revising familiar modes of research and communication in a seemingly foreign language (Box 2). Yet hewing to traditional practices carries its own risks for the individual investigator. Errors and oversights can persist far longer when experimental design, raw data, and data analysis are held in private; even once published, weeks and months can be wasted in chasing reproduction of results because methods are documented only as fully as a journal word count permits; labs can become isolated, their advancement slowed, for lack of substantive interaction with others. As has been demonstrated in other disciplines, open science can help to mitigate these risks, to the immediate benefit of the individual practitioner (Davis and Fromerth 2007; Lawrence 2001).

Moreover, open science promises many longer term benefits to the scientific community. The adoption of standard best practices and cultural norms for public archiving of data and code will advance discovery and promote fairness in attribution. The use of open-source tools and open-access data and journals will help to further democratize science, diversifying perspectives

PeerJ PrePrints

and knowledge by promoting broader access for scientists in developing countries and at under-resourced institutions, fostering the citizen science that is already a major source of data in some ecological sub-disciplines (Cooper *et al.* 2014), and improving the communication of scientific findings to the general public (Fausto *et al*. 2012).

Here, we discuss the changes in mindset and the tools that can help interested ecologists to move toward practicing open science themselves, to facilitate its practice by their students and other colleagues, or both.

Changes in mindset

Data stewardship, not data ownership

Traditional views on data ownership hold that data are proprietary products of the researcher (Sieber 1989). By definition, this data-ownership mindset limits the potential for data sharing as a given researcher can restrict the conditions and circumstances by which their data are disseminated. These views have persisted for a variety of reasons (Sieber 1989, Hampton *et al.* 2013, Lindenmeyer and Likens 2013) and ecologists historically have treated data as proprietary, whether or not the data collection has been funded by taxpayers and might reasonably be considered public property (Obama 2013).

Under the principles of open science, data are generated with the expectation of unfettered public dissemination. This fundamental shift in thinking from "I own the data" to "I collect and share the data on behalf of the scientific community" is essential to the transparency and reproducibility of the open science framework. When data are available, discoverable, reproducible, and well-described, scientists can avoid “reinventing the wheel” and instead build

PeerJ PrePrints

directly on those products to innovate. For example, authors' reluctance to submit null results for publication leads to a "file-drawer" effect that can not only systematically bias the published literature (Iyengar and Greenhouse 1988, Franco *et al.* 2014), but also allows independent scientists to go repeatedly down the same blind alleys. Structures to store, share, and integrate data contribute to preventing such waste of scientific and public resources. Beyond this greater efficiency, data sharing also contributes to the production of entirely new scientific products that were not envisioned at the time data were collected (Carpenter *et al.* 2009).

Norms have yet to be established in ecology for how soon, after collection, data should be shared in order to promote openness and a healthy scientific culture. Indeed, scientists who are philosophically aligned with open science currently employ a range of data sharing practices (Figure 2). A full embrace of open science implies sharing data instantaneously, or upon completion of initial quality assurance checks or other pre-processing (e.g., NEON; LTER Taylor and Loescher 2013). In other cases, researchers have made an argument for a constrained period of exclusive access by researchers directly involved in data collection (e.g., Sloan Digital Sky Survey; <http://www.sdss.org/>). Despite these differences, it is increasingly recognized in the requirements of funding agencies that full data sharing in established repositories should begin no later than the publication of results.

Transparency throughout the data life-cycle

Scientists publish their methodology as a means to enable reproducibility by others, but have traditionally had to judge which details were important to transmit within the limitations imposed by print journals. The availability of online supplementary methods sections gives

PeerJ PrePrints

scientists scope to detail their methods more fully, and a broader suite of online tools now creates opportunity to share the code, data, and detailed decision making processes that constitute the scientific endeavor. Taking advantage of these opportunities to make tacit knowledge explicit to others is a crucial part of performing reproducible science (Collins 2001, Ellison 2010), and provides the substantial additional benefit of exposing untested assumptions and unidentified confounding effects.

Workflow tools (Table 1) now make it possible for scientists to make nearly every stage of the research process transparent, from sharing the detailed rationale for an approach to publishing the data and code that generated analyses and figures. Detailed sharing of methods and code improves clarity; personal communications regarding methods crucially improves trust (Collins 2001), and social media permit these communications to happen in the open. Openness throughout the data life-cycle also provides the scientist with the opportunity to receive feedback from the rest of the scientific community and the general public, reducing redundancy and accelerating scientific inquiry (Byrnes *et al*. 2014). Whereas formal peer review provides feedback only at the project's proposal phase (for those seeking grant support) and at publication, open science provides an avenue for scientists to receive feedback at key junctures, *e.g.*, before experiments are performed.

Additionally, transparency encourages researchers to converge on standard structures for data and code archiving (Table 1). Such convergence is particularly important for interdisciplinary science, in which the fragmentation of resources and practices along disciplinary boundaries can substantially hinder research. Common standards and a shared,

PeerJ PrePrints

searchable infrastructure help make data sets not merely open but also discoverable, improving their reach and impact and helping scientists identify potential new collaborators.

Having said all this, scientists need not fear that open science is only for the exhibitionists among us; we recognize that there are many points in the scientific process when deep, sometimes solitary reflection is invigorating and productive.

Acceptance of critique

Failure is recognized as a normal and necessary part of the scientific process, and yet academic science is structured to reward only being right in public (Merton 1957), creating tension in practicing open science. The more open our science, the greater the chance that our mistakes as well as our insights will be public. This prospect can be frightening to contemplate; one study of physicists found that those practicing secrecy prior to publication often did so to avoid the risk of looking foolish (Gaston 1971). We suggest that embracing this tension gives us the opportunity to be better and more productive scientists. The only way to protect our ideas and methods from criticism indefinitely is to refrain from publication, hardly a desirable outcome. Even delaying exposure until peer review or post-publication (Sabine 1985) manages only to limit the possible range of feedback to, essentially, “what could have been done better”. By contrast, adopting open practices throughout the scientific endeavor makes it possible to receive and incorporate critiques before our research products are complete. That is, by risking the possibility of being briefly wrong in public, we improve our chances of being lastingly, usefully right.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

Tools and best practices to enable shifts in mindset and practice

An open science mindset affects the entire scientific process, carrying responsibilities and offering benefits at each stage along the way (Figure 2). Throughout the process, social media are used to publicly discuss ideas, hypotheses, experimental designs, data collection, analytical approaches, and eventually publication (Gewin 2013, Darling *et al*. 2013). Products are published in open repositories that provide stable identifiers, version control and time stamps (Noble 2009, Wilson *et al*. 2014). Version control systems allow scientists to retain snapshots of previous analyses for future reference, collaborate easily and track contributions, record ideas, and safeguard against the loss of code and data (Ram 2013), thus preserving the long-term integrity of the project even as collaborations form and shift. Stable identifiers (URL or DOI) for every product allow proper attribution and linking. All of these steps are undertaken with an eye to making our work reproducible and open to others, but all offer the immediate benefit of making our work reproducible to ourselves. Many of the tools mentioned in Table 1 have proprietary analogs (*e.g.*, as SAS is to R), and afford many similar advantages, but exclusive use of open-source, free software maximizes access by other researchers. All of these tools give us access to a research group far bigger than a single lab, helping experimental designs to be improved and stimulating discussion of worthwhile new directions, connections, and approaches.

Data: If we are committed to data stewardship, planning an experiment entails not only thinking through the physical manipulations involved but also working out how to capture and share the data and metadata that will enable others to effectively re-use that information. The open-source DMPTool (Table 1) offers guidance to scientists creating data management plans—now often a prerequisite for funding—and helps scientists find institutional resources for

PeerJ PrePrints

implementation. At the same time, ready access to data sets collected by other scientists can help focus our questions, by identifying gaps and opportunities, and improve our ability to answer them (*e.g.,* by allowing us to estimate and plan for experimental uncertainties). Once data have been collected, the open scientist prepares the data set for use by others and documents its provenance, then deposits it in a community-endorsed repository (e.g., Knowledge Network for Biocomplexity, Dryad) (Rüegg *et al*. 2014). Many software tools facilitate the sharing and documentation of data as well. Tools created by the ROpenSci initiative allow integration of this process within R based workflows with packages such as EML (Metadata creation) and rfigshare (data sharing on figshare.com). User-friendly tools such as Kepler or VisTrails help document provenance, and Morpho is an easy way to create EML (standardized, machine-readable) metadata. This process ensures that the data will remain usable, accessible, and become citable for years to come. It further allows our work to be recognized and integrated more effectively into the larger body of knowledge and ensures that, when we return to a project after a period away, we can pick up where we left off.

Research Process: If we are committed to transparency, we document and share as much information about the actual research process as is feasible. Electronic lab notebooks (*e.g.,* using IPython notebooks) help track and share the reasoning behind our experimental and analytical decisions, as well as the final protocol and any deviations, and can be linked to the resulting data files to keep research organized. Adhering to the discipline of consistently, carefully, and thoroughly documenting the research process is an exercise in critical thinking, a constant reminder to check our assumptions and clarify our thinking.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

PeerJ PrePrints

Data Analysis: During data analysis, reproducible, script-based methods (*e.g.,* in R or Python) can be used for every step from importing raw data to analysis and production of figures and final manuscripts (*e.g.,* FitzJohn et al 2014). Such tools are essentially self-documenting along the way. However, they may still produce many separate scripts, which would have to be executed in sequence. Workflow systems, like Kepler or VisTrails, can provide a more complete record of data manipulations. This type of record is almost impossible to generate for point-and-click analyses in a graphical user interface (GUI). While errors can be made in both scripted and GUI-based analyses, the existence of a record makes errors in the former far easier to detect and correct, protecting the integrity of the analysis. Tools such as Sweave and knitr facilitate integration of data analysis into manuscript production, making it easier to keep figures and reported results current as an analysis is refined.

Publication: Presentations, posters, figures, and movies can be opened for comment on public websites (*e.g.,* Figshare, SlideShare). Publication preprints can be posted for comment from an audience broader than a journal's handful of peer reviewers; preprints also improve a project's visibility and, with the addition of a date stamp, establish precedence (Desjardins-Proulx *et al.* 2013). Publishing final papers in open-access journals ("gold" open access) or self-archiving manuscripts ("green" open access) makes the final products available to a wide audience, including the taxpayers who may have funded the research.

Conclusions

Online tools make possible a future in which not only scientific practice but also scientific culture is transformed by openness (Nielsen 2011). Fully open science can take place

PeerJ PrePrints

throughout the process of discovery, from the sharing of nascent ideas, to the uploading of data at the moment of capture, through to developing "living papers" in an open forum in which the details of analysis and reasoning are completely transparent. Subsequent generations of ecologists will build their work on what we leave. If instead of exclusive silos of traditional journal archives, we leave future researchers open-access repositories of data, code, and papers, they will be far better equipped to push new frontiers in science and create solutions to pressing societal problems.

Very real technological and cultural hurdles still stand between us and this future: investigators must be willing to invest time in learning the tools that facilitate open science, and in re-learning them as the tools evolve. Further, the scientific community must collectively establish new norms for collegiality and reproducibility in the digital age. Nevertheless, we can all move our research toward this future by adopting the aspects of open science that are currently feasible for our own research groups (*e.g.,* publishing open-access articles; sharing all data and code used in publications) and by supporting our students and junior colleagues in developing the skills that will best prepare them for the responsibilities, opportunities, and rewards of practicing ecology in an open environment.

Acknowledgments

Much of the work on this manuscript was done at the Open Science Codefest which was supported by awards from the National Science Foundation (1216894 and 1216817) and sponsored by the Renaissance Computing Institute, the National Center for Ecological Analysis and Synthesis, rOpenSci, DataONE, and Mozilla Science Lab. M. Cantiello, E. Robinson, L.

PeerJ PrePrints

Winslow, J. Couture, C. Granade, S. Earl, J. Ranganathan, and D. LeBauer provided valuable ideas and feedback.

Literature Cited

Beaver, D. deB, and R. Rosen. 1978. Studies in scientific collaboration. Scientometrics 1:65–84.

Byrnes, J. E. K., E. B. Baskerville, B. Caron, C. Neylon, C. Tenopir, M. Schildhauer, A. Budden, L. Aarssen, and C. Lortie. 2014. The four pillars of scholarly publishing: The future and a foundation. Ideas in Ecology and Evolution 7.

Carpenter, S. R., E. V. Armbrust, P. W. Arzberger, F. S. Chapin, J. J. Elser, E. J. Hackett, A. R. Ives, P. M. Kareiva, M. A. Leibold, P. Lundberg, M. Mangel, N. Merchant, W. W. Murdoch, M. A. Palmer, D. P. C. Peters, S. T. A. Pickett, K. K. Smith, D. H. Wall, and A. S. Zimmerman. 2009. Accelerate Synthesis in Ecology and Environmental Sciences. BioScience 59:699–701.

Casadevall, A., and F. C. Fang. 2010. Reproducible Science. Infection and Immunity 78:4972–4975.

Cerf, V. 2002. The Internet is for Everyone. http://tools.ietf.org/html/rfc3271.

Chan, L., D. Cuplinskas, M. Eisen, F. Friend, Y. Genova, J.-C. Guédon, M. Hagemann, S. Harnad, R. Johnson, R. Kupryte, M. La Manna, I. Rév, M. Segbert, S. de Souza, P. Suber, and J. Velterop. 2002, February 14. Budapest Open Access Initiative. http://www.opensocietyfoundations.org/openaccess/read.

Collins, H. M. 2001. Tacit Knowledge, Trust and the Q of Sapphire. Social Studies of Science 31:71–85.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

Cooper, C. B., J. Shirk, and B. Zuckerberg. 2014. The Invisible Prevalence of Citizen Science in Global Research: Migratory Birds and Climate Change. PLoS ONE 9:e106508.

Darling, E.S., D. Shiffman, I.M. Côté, J.A. Drew. 2013 The role of Twitter in the life cycle of a scientific publication. Ideas in Ecology and Evolution 6: 32 – 43, 2013 doi: 10.4033/iee.2013.6.6.f

Davis, P.M., M.J. Fromerth 2007. Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? Scientometrics 71: 203-215

Desjardins-Proulx, P., E. P. White, J. J. Adamson, K. Ram, T. Poisot, and D. Gravel. 2013. The Case for Open Preprints in Biology. PLoS Biol 11:e1001563.

Dietze, M. C., D. S. Lebauer, and R. Kooper. 2013. On improving the communication between models and data. Plant, Cell & Environment 36:1575–1585.

Drummond, D. C. 2009. Replicability is not Reproducibility: Nor is it Good Science. Proceedings of the Evaluation Methods for Machine Learning Workshop 26th ICML, Montreal, Quebec, Canada. http://www.csi.uottawa.ca/~cdrummon/pubs/ICMLws09.pdf.

Ellison, A. M. 2010. Repeatability and transparency in ecological research. Ecology 91:2536–2539.

Fausto S, Machado FA, Bento LFJ, Iamarino A, Nahas TR, et al. (2012) Research Blogging: Indexing and Registering the Change in Science 2.0. PLoS ONE 7(12): e50109. doi:10.1371/journal.pone.0050109

FitzJohn, R. G., M. W. Pennell, A. E. Zanne, P. F. Stevens, D. C. Tank, and W. K. Cornwell. 2014. How much of the world is woody? Journal of Ecology 102:1266–1272.

Franco, A., N. Malhotra, and G. Simonovits. 2014. Publication bias in the social sciences:

PeerJ PrePrints

Unlocking the file drawer. Science:1255484.

Gacek, C., and B. Arief. 2004. The many meanings of open source. IEEE Software 21:34–40.

Gaston, J. 1971. Secretiveness and competition for priority of discovery in Physics. Minerva 9:472–492.

Gewin, V. 2013. Turning point: Carl Boettiger. Nature 493:711–711.

Hackett, E. J. 2005. Essential Tensions Identity, Control, and Risk in Research. Social Studies of Science 35:787–826.

Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter. 2013. Big data and the future for ecology. Frontiers In Ecology And The Environment 11:156–162.

Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. Nature, 482(7386), 485–8. doi:10.1038/nature10836

Iyengar, S., and J. B. Greenhouse. 1988. Selection Models and the File Drawer Problem. Statistical Science 3:109–117.

Jasny, B. R., G. Chin, L. Chong, and S. Vignieri. 2011. Again, and Again, and Again …. Science 334:1225–1225.

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. Nature, 411, 521. doi:10.1038/35079151

Lindenmayer, D., and G. E. Likens. 2013. Benchmarking Open Access Science Against Good Science. Bulletin of the Ecological Society of America 94:338–340.

Merton, R. K. 1957. Priorities in Scientific Discovery: A Chapter in the Sociology of Science. American Sociological Review 22:635–659.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

PeerJ PrePrints

Michener, W. K., and M. B. Jones. 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology & Evolution 27:85–93.

Molloy, J. C. 2011. The Open Knowledge Foundation: Open Data Means Better Science. PLoS Biol 9:e1001195.

Mosley, M., M. H. Brackett, S. Earley, and D. Henderson. 2009. DAMA guide to the data management body of knowledge. Technics Publications.

National Science Foundation. 2007. NSF and the Birth of the Internet. http://www.nsf.gov/news/special_reports/nsf-net/.

Nielsen, M. 2011. Reinventing Discovery: The New Era of Networked Science. Reprint edition. Princeton University Press.

Noble, W. S. 2009. A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5:e1000424.

Obama, B. 2013, May 9. Executive Order -- Making Open and Machine Readable the New Default for Government Information. The White House.

Peng, R. D. 2009. Reproducible research and. Biostatistics 10:405–408.

Ram, K. 2013. Git can facilitate greater reproducibility and increased transparency in science. Source Code for Biology and Medicine 8:7.

Reichman, O. J., M. B. Jones, and M. P. Schildhauer. 2011. Challenges and Opportunities of Open Data in Ecology. Science 331:703 –705.

Rüegg, J. C. Gries, B. Bond-Lamberty, G. J. Bowen, B. S. Felzer, N. E. McIntyre, P. A. Soranno, K. L. Vanderbilt, and K. C. Weathers (2014). Closing the data life cycle: Using information management in macrosystems ecology research. Frontiers in Ecology and the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

Environment 12(1):24-30. doi:10.1890/120375.

Sabine, J. R. (1985). The Error Rate in Biological Publication: A Preliminary Survey. BioScience, 35(6), 358–363. Retrieved from http://www.jstor.org/stable/1309904

Sieber, J. E. 1989. Sharing Scientific Data I: New Problems for IRBs. IRB: Ethics and Human Research 11:4.

Stodden, V. 2009. Enabling Reproducible Research: Open Licensing for Scientific Innovation. SSRN Scholarly Paper, Social Science Research Network, Rochester, NY.

Stodden, V., J. Borwein, and D. H. Bailey. 2013. "Setting the default to reproducible" in computational science research. SIAM News 46:4–6.

Strasser, C. A., and S. E. Hampton. 2012. The fractured lab notebook: undergraduates and ecological data management training in the United States. Ecosphere 3:art116.

Taylor, Jeff R, and Henry L Loescher. (2013) "Automated Quality Control Methods for Sensor Data: A Novel Observatory Approach." *Biogeosciences* 10(7): 4957–71.

Wagner, C. S. 2008. The New Invisible College: Science for Development. Brookings Institution Press.

Wilson, G., D. Aruliah, C. T. Brown, N. P. C. Hong, M. Davis, R. T. Guy, S. H. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, and others. 2014. Best practices for scientific computing. PLoS biology 12:e1001745.

Winter, D and Chamberlain, S 2014, rentrez: Entrez in R version 0.3.1 http://cran.r-project.org/web/packages/rentrez/index.html

Wuchty, S., B. F. Jones, and B. Uzzi. 2007. The Increasing Dominance of Teams in Production of Knowledge. Science 316:1036–1039.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

Table 1: A wide range of tools is available to support open science at each stage of the research life-cycle. Tools change over time and these are some, but not all available, at the time of this publication.

Concept	**Name of Tool or Service**	**Tool Description**
Ideas & communication		
Open discussion	Twitter	Twitter allows users to write, share, and respond to short 140-character messages. An ever-increasing community of scientists uses Twitter to share ideas about research (Darling et al. 2013).
	Blogs	Blogs can be hosted on university websites, personal servers or blogging sites (e.g., wordpress.com). Blogs offer an informal means of discussing ideas, results, published literature, etc.
	Open lab notebooks	Open lab notebooks apply the concept of blogging to day-to-day research work: research notes and data are published online as they are accumulated.
	GitHub comments	GitHub comments allow others to review code through its development and offer comments on particular sections.
Technical support	StackOverflow	StackOverflow is a general question and answer site for programming problems, supplementing the help sites available for most programming and scripting languages.
Hypotheses/Design	Data Management Planning Tool	The Data Management Planning Tool enables researchers to easily create, manage and share data management plans that meet the requirements of a broad array of funding agencies and institutions.
Data Life-cycle Support	Data repositories: KNB, DataONE, Dryad, GBIF	Data repositories make data available to future researchers and allow research to be reproduced; they are a cornerstone of open science.
	Open Office	Open Office is a comprehensive open-source office tool suite that supports word processing, spreadsheets, graphics, presentations, drawing, and creating and maintaining databases.
	MySQL	mySQL is a popular and widely used open-source relational database management system (RDBMS)

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

PeerJ PrePrints

		based on Structured Query Language (SQL).
	OpenRefine	Web-based tools for working with data.
	Morpho	Morpho is a program that can be used to enter metadata, which are stored in a file that conforms to the Ecological Metadata Language (EML) specification.
Analysis and Visualization		
Version control	Git and GitHub	Git is a piece of software that allows you to create 'versions' of your code, text, and project files as you work on them. GitHub is a website that allows this to be done collaboratively, with social and discussion features built in.
Visualization of geospatial data	GRASS	GRASS (Geographic Resources Analysis Support System), is a Geographic Information System (GIS) software toolset used for geospatial data management, analysis, and visualization, as well as image processing and spatial modeling.
	QGIS	QGIS is a desktop GIS application that supports geospatial data viewing, editing, and analysis.
Workflow tools	Kepler	Kepler is a scientific workflow package that allows researchers to create, execute, and share analytical models.
	VisTrails	VisTrails is a scientific workflow and provenance management system that supports data exploration and visualization.
Reproducibility	R	R is a widely used statistical programming language that is commonly used for analyzing and visualizing data.
	RStudio	RStudio is an Integrated Development Environment (IDE) for R.
	Python	Python is a widely used high -evel programming language that is commonly used for managing and manipulating data.
	Pycharm	Pycharm is one of several IDEs available for python.
	IPython notebook / Project Jupyter	The IPython notebook (now renamed Project Jupyter and focusing on R and Julia in addition to python) is a tool for interactively analyzing and processing data in the browser using blocks of code.
	Sweave	Sweave was originally a way to integrate S and LaTeX, but now also works with R.

PeerJ PrePrints

	markdown	Markdown is a simple markup syntax for adding formatting to documents. It allows correctly formatted scientific documents to be written in plain text.
	pandoc	Pandoc allows conversion between many document types, including LaTeX, markdown, PDF, and Word (.docx).
	knitr, Babel	knitr (originally for R) and Babel (an Emacs extension) allow the integration plain narrative text with blocks of code in many different scripting languages within a single document.
	Rmarkdown	Rmarkdown is an authoring format which combines markdown with the syntax of both knitr and pandoc.
Writing		
Collaboration	Google Docs	Google Docs is a suite of online collaborative writing, spreadsheet, and presentations tools.
	Etherpad	Etherpad is an online, open source, collaborative writing tool.
	ShareLateX, WriteLaTeX, Authorea	These are online collaborative writing tools focused on LaTeX.
Reference management	Zotero	Zotero is a free and open-source extension to the Firefox browser (and now a standalone app) for literature management and citation.
	Mendeley	Mendeley is a free reference manager and social network for researchers.
Presenting Preliminary Results		
Distribution of figures and talks	Figshare	Figshare is an online repository for all types of research products (data, posters, slides, etc) that assigns each a citable DOI.
	Slideshare	Slideshare is an online clearinghouse for presentation slides of all types.
	Speakerdeck	Speakerdeck is an online site, run by Github, for sharing PDF presentations.
Distribution of preprints	bioRXiv	bioRXiv, run by Cold Spring Harbor, is a relatively new preprint server that focuses primarily on biological research.
	arXiv	arXiv is one of the original preprint servers on the web. Run by Cornell, it is mainly focused on math, physics, and computer science, although it has been used by

		quantitative biologists as well.
	PeerJ Preprints	PeerJ Preprints is a preprint server run by the open-access online-only journal PeerJ.
Pre-publication peer preview	Peerage of Science	Peerage of Science offers pre-publication formal peer review (and review of the reviews), which can then be sent on to participating journals.
	Axios Review	Axios Review offers pre-publication formal peer review and appraisal of a manuscript's fit with targeted journals; reviews can then be sent on to participating journals.
Publication	DOI for code	Code can be given a DOI and cited in the literature. For example, a Github repository can be assigned a DOI via zenodo.org.
	DOI for data	Data uploaded to any of the numerous available online repositories will be assigned a DOI and is then citable by other researchers using that dataset.
	"Green" open access	"Green" open access is the posting of a research article pdf to an author's personal website.
	"Gold" open access	"Gold" open access is the open publication of a paper on the journal website, funded by an up-front (pre-publication) fee paid by the authors.
	Licences: CC-BY, CC-BY-NC etc	Licenses dictate how a research product may be used by others (e.g., requiring attribution or prohibiting commercial reuse).
Discussion of published literature and data		
Discovery of published data	DataONE	DataONE is a federation of data repositories that supports easy discovery of and access to environmental and Earth science data, as well as various data management tools and educational resources.
	re3data	re3data is a registry of digital repositories that enables researchers to discover public and institutional repositories where they may deposit and preserve their data.
Social networking	ResearchGate	ResearchGate is a social networking and question and answer site to which researchers can also upload their publications.
	Academia.edu	Academia.edu is a social network for academics.

PeerJ PrePrints

Tracking research product impact	ORCID	ORCID provides unique identifiers for individual researchers, which allows contributions to be tracked across many repositories, grant proposals, peer review sites, etc.
	ImpactStory	ImpactStory can track almost all of the research contributions (data, code and papers) by individual researchers, and quantifies their impacts using open data sources (e.g., tweets, use in wikipedia articles, saves in Mendeley).
	Altmetric	Provides metrics (tweets, blog posts, Mendeley saves, etc) of individual research objects.
Informal discussion	Conference or hallway conversations, discussion groups	These conversations are highly efficient but offer limited accessibility to outside researchers.
	Personal website/blog	Personal blogs can be a forum to discuss both one's own research as well as the research of other scientists.

PeerJ PrePrints

Box 1: Technological advances driven by scientists

Every scientist now uses the Internet, but few are aware of how the Internet grew out of a highly collaborative and open process involving development of publicly available and commentable standard protocols (<http://www.fcc.gov/openinternet>; Cerf 2002). The availability of "open source" software (a term first coined in the 1990s) radically democratized and expanded participation in the Internet community in the late 1980s-early 1990s. "Open source" encompasses not only compilers and applications but also protocols and specifications such as the domain name system (DNS) that allows pinpointing specific networked computers ("hosts") around the world, and HTTP/HTML specifications that provide the basis for the World Wide Web.

Members of the scientific research community were early recipients of these advantages, with the National Science Foundation supporting and nurturing growth of the Internet-based NSFNET from roughly 1985-1995 (National Science Foundation, 2007). In that era, it was scientists who were largely communicating through the Internet (gopher, email), transferring their data (FTP), and running analyses on remote servers (telnet, shell access, X11), often with privileged access to fast networks and accounts on powerful computational servers. Within this computer savvy community, "power users" leveraged the Internet most effectively via learning computational skills that were largely command-line based. The legendary, free GNU suite of software was standard issue for many computers joining the Internet in the late 1980s, and made that early generation of networked "scientific workstations" (from Sun, SGI, DEC, or NeXT) the sought-after systems of their day.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

These early forays into powerful software helped birth the plethora of tools now available to the modern scientist. Today, free, multi-platform, open source tools from the Linux Foundation (free operating system), the Apache Software Foundation (free Web server), the Mozilla Foundation (free Web, email, and other applications), the PostgreSQL Global Development Group (free enterprise database), the Python Software Foundation (free programming language), and the R Foundation for Statistical Computing (analysis and statistical language) are enabling researchers across the globe to dialog with one another via cutting edge communication, execute powerful data manipulation, and develop community-vetted modeling and analysis tools at minimal individual cost.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

Box 2: "A glossary of open science for ecologists"

citizen science: enabling interested citizens to contribute their time, observations, and expertise to assist and inform the scientific research process; may be an aspect of ***crowd-sourcing***.

code repository: an accessible, central place where computer code is stored to facilitate the collection, manipulation, analysis, or display of data.

crowd-sourcing: leveraging the expertise and participation of many individuals, to provide more perspectives, critiques, data contributions, code contributions, etc. to advance a (scientific) process.

data life-cycle: the pathway researchers trace when confronting a challenge with data, from idea generation through to making observations and drawing inference. Popularly dissected into eight intergrading phases: Plan, Collect, Assure, Describe, Preserve, Discover, Integrate, Analyze (Michener and Jones 2012).

data management: the development and execution of architectures, policies, practices and procedures that properly manage the full ***data life-cycle*** needs of an enterprise (Mosley *et al.* 2009).

data repository: an accessible, central place where accumulated files containing collected information are permanently stored; typically these house multiple sets of databases and/or files.

open access: providing free and unrestricted access to research products, especially journal articles and white papers—to be read, downloaded, distributed, reanalyzed, or used for any other legal purpose—while affording authors control over the integrity of their work and the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

PeerJ PrePrints

right to be acknowledged and cited (adapted from the Budapest Open Access Initiative definition, Chan *et al.* 2002).

open data: data that can be freely used, reused, and redistributed without restrictions beyond a requirement for attribution and share-alike (Molloy 2011).

open source: computer code (software) that is available for free distribution and re-use, with source code unobscured, and explicit acknowledgement of the right to create derived works by modifying the code (Gacek and Arief 2004)

open science: the idea that scientific knowledge—including data, observational and experimental design and methods, analytical and modeling code, as well as results and interpretations of these (e.g., as reported in publications)—can and should be made freely accessible to anyone, and represented in transparent and reusable formats as early as practical in the discovery process, by employing standards-based technology tools. Frequently encompasses all of ***open access***, ***open data*** and ***open source*** and, minimally, facilitates reproducibility of results.

preprint: a draft version of a paper distributed (usually in an online repository such as arXiv) before a final, peer-reviewed journal or reporting agency has accepted or formally published the paper (Desjardins-Proulx *et al.* 2013).

provenance: the origin of data, including any transformations occurring along the way

reproducibility, replicability, and repeatability: while formal definitions of these terms vary widely and across disciplines, these all point to a hallmark of science, which is the ability to repeatedly generate or observe outcomes consistent with scientific understanding, based on explicit specification of theories, models, and methods, and their expected material realizations

PeerJ PrePrints

or outcomes. This concept prescribes a need for sufficient access to data and analytical code to verify that a purported result is valid, as well as to examine these for errors and biases (Jasny et al 2011; Peng 2009; Stodden et al. 2013; Stodden 2009; but note Drummond 2009 and Casadevall and Fang 2010 use somewhat different definitions)

transparency: sufficiently detailed description of a scientific process to enable meaningful public scrutiny and examination, with nothing intentionally obscured by technology or process

version control: a system that manages snapshots (and hence "revisions" or "versioning") of code and data for a project (Wilson *et al.* 2014). Facilitates detailed documentation to enable tracing any significant changes over a project's lifetime.

PeerJ PrePrints

PeerJ PrePrints

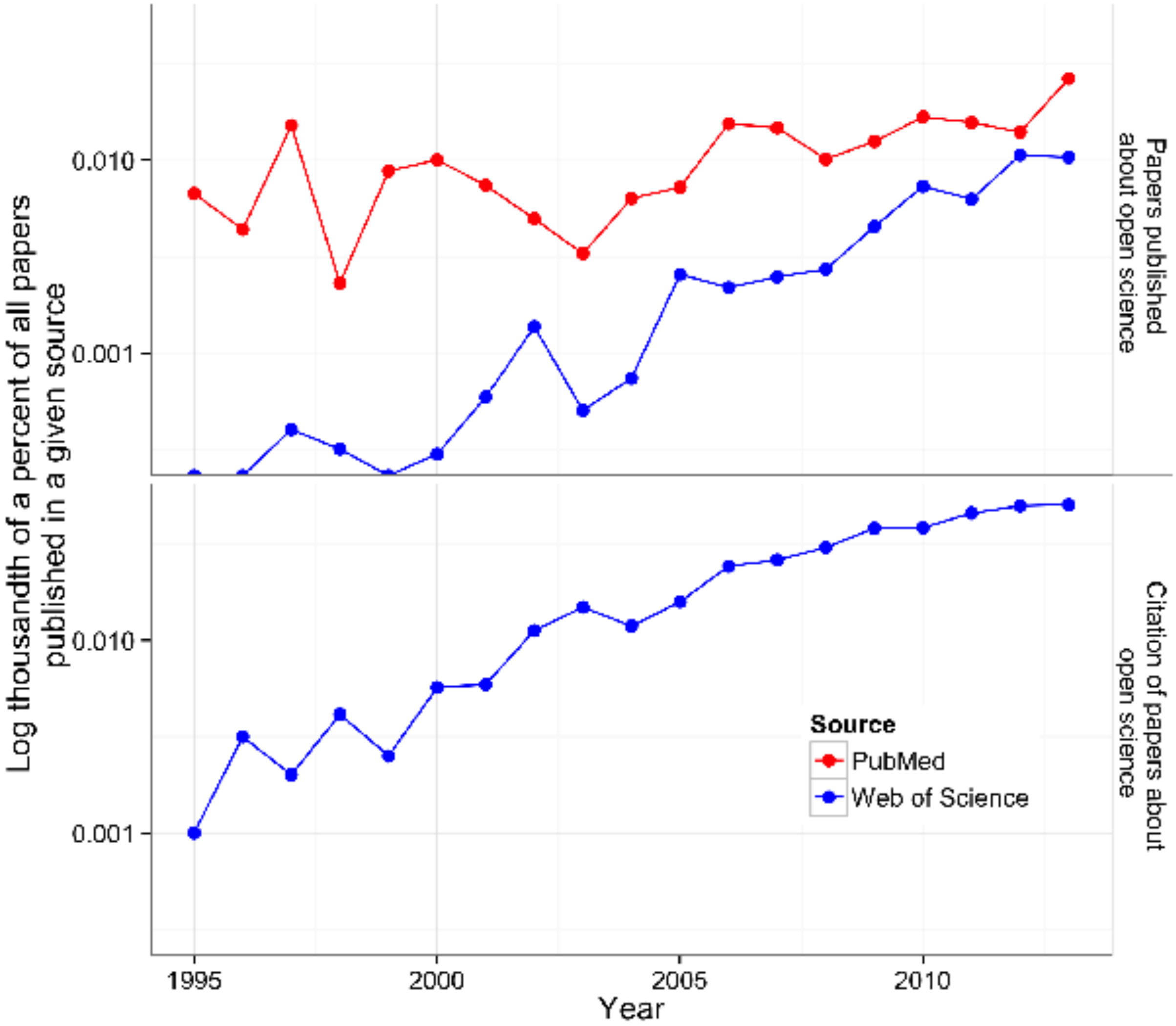

Figure 1. Increasing usage of the term "open science" in the literature since 1995 in Web of Science and PubMed databases. Data from PubMed were downloaded via the rentrez (Winter and Chamberlain 2014) package in R, and Web of Science data were collected from manual searches. Results were normalized by total articles published each year to account for the increasing number of publications. Both data sources show an increase in the number of publications about open science, and an increase in annual citations of those papers.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.549v1 | CC-BY 4.0 Open Access | rec: 21 Oct 2014, publ: 21 Oct 2014

PeerJ PrePrints

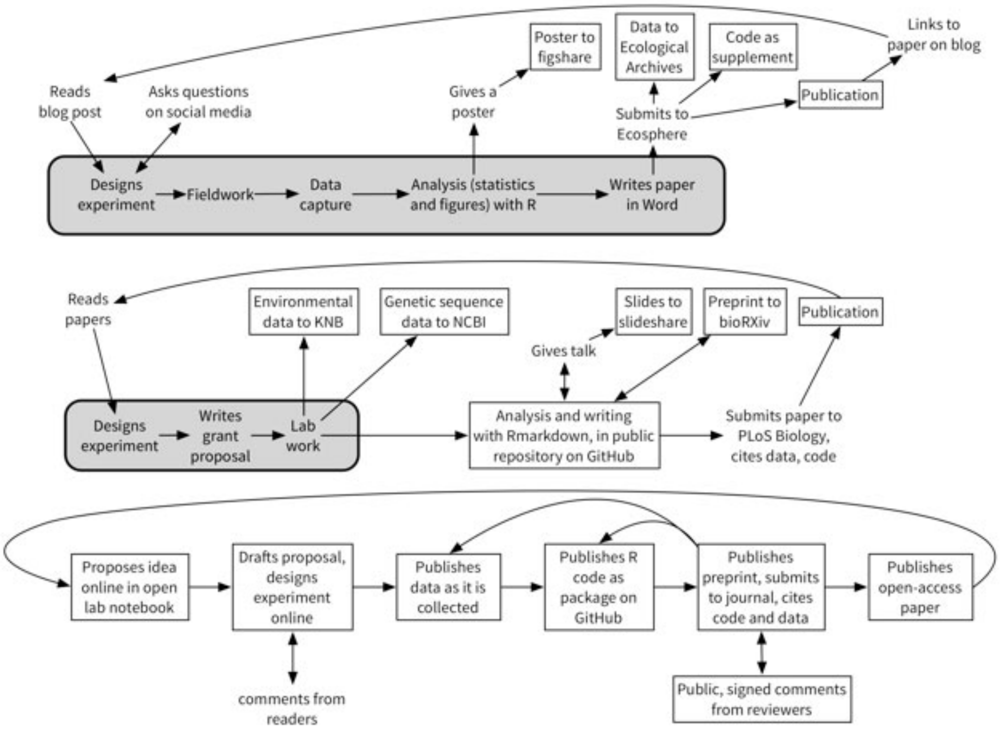

Figure 2: Three examples of possible open science workflows. In each workflow illustration, the *gray box* surrounds activities that are not openly accessible for researchers who are not directly involved. Activities outside these boxes are open and available, representing how the individual researcher is influenced by other scholars, or is able to communicate their research before and after publication. *White boxes* represent distinct research products available for reference for and feedback from other researchers.