Open Sourcing Ecological Data

Viewpoint article

Cynthia Sims Parr

Adjunct Professor, Behavior Ecology Evolution and Systematics
Research Associate, Univ. Maryland Inst. Adv. Computer Studies
A.V. Williams Building
University of Maryland
College Park, MD 20742

and

Assistant Research Professor, Computer Science and Electrical Engineering
University of Maryland Baltimore County
1000 Hilltop Cir.
Baltimore, MD
410-455-8938

csparr@umd.edu

In a thought-provoking viewpoint, Cassey and Blackburn (2006) suggest that reproducibility should not be required of ecological studies. Thus, ecological journals should not require authors to publish data as a requirement of publication, nor should reviewers insist on it. They make three cautionary points: First, the goal of reproducibility should not be applied piecemeal. Second, journals are not ready for custodianship of data. Third, publishing data places the intellectual rights of authors at risk under the current reward system. I will respond to each of these points, then end with another view of the future of ecological research: an open source web of ecological data.

Reproducibility and Re-use

I agree that a reproducibility requirement should not be applied indiscriminately. Cassey and Blackburn expect scientific fraud, data loss, or error to be evenly distributed across research areas. However, reproducibility can be more important in some areas than in others. Transparency and verifiability in politically sensitive research areas (global climate change, conservation biology, etc.) are in the best interests of scientists and society. Controversy over phylogenetic reconstruction methods has led to the expectation that character data be published so it can be re-analyzed. Conflicting studies that bear on human health are subjected to meta-analyses using pooled data if available. Research bearing on ecosystem health should be treated similarly.

Even if published data are not *necessary* for evaluation of a specific study or research question, potential value of data to the community can be a *sufficient* reason to require its publication. Ecologists should not be subject to a lower standard than life scientists in genomics and medicine. Certainly novel uses of data have already advanced the science

of ecology, as synthetic studies increasingly produce knowledge on scales not previously possible.

Recommendations: Dialogue within the scientific community is necessary to help journals and reviewers determine when reproducibility and re-usability are most desirable. Then data-sharing requirements can be consistently and fairly applied. The NRC and the Ecological Society of America recommend broad data sharing (NRC 2003, Palmer et al., 2004). However, all journals need not come to the same conclusions. If top tier journals choose to have stricter requirements than other journals, this should be a factor in whether one chooses to submit to them.

Journals as data custodians

The authors express concern that journals are not ready to be custodians of original data. Larger publishers already archive data and supplemental material (e.g. *Ecology, Science, Nature*) largely in flat formats; such data may not be the most easily found or used, but there is significant value in its likely longevity. Still, journals need not be data custodians, nor is a universal protocol, framework, or intellectual property policy necessary. Other long term repositories and registries are maintained by universities, government agencies, and other institutions which are already working with the scientific community to develop standards and protocols (reviewed in Parr and Cummings 2005, Jones et al. 2006). The best way to speed progress on these is to use the one best suited for your needs, not to wait until they are perfect. When these choices are coupled with a changed reward system described below, the most effective standards, policies, and protocols can emerge in a darwinian process. Just as there is no universally successful

suite of adaptations, there will never be a perfect set of standards, protocols, and software to be applied to all science.

Recommendations: Each journal should create a consistent policy about where and how and when data associated with their publications must be archived. Until clear community standards emerge, journals should provide a wide range of choices from which authors can choose. Authors and their institutions can therefore "vote with their data" (and indeed, their manuscripts) and begin to influence which repositories, standards, tools, and policies become the community standards.

A changing scientific reward system

Cassey and Blackburn are understandably concerned about the impact of data sharing on the rights of authors. It would seem that openly providing data would leave one at risk of being scooped, left out of collaborations, or unrewarded by other researchers.

Data-driven studies may seem easy but new fields of bioinformatics, ecological informatics, and biodiversity informatics have sprung up with specialized skills and training. These skills are necessary to effectively integrate and re-use the data, a process fraught with pitfalls (Blackburn and Gaston 1998, Jones et al. 2006). If others would put your dataset to the same future use you plan for it, aren't you as the originator of the data going to have a head start and be more likely to do the better job?

Informatics studies are often highly collaborative. Most likely the best people for such collaborations are the ones who have the deepest understanding of how data were collected and previously analyzed. Given a choice between someone who has already made data available in formats you can see are likely to be useful, and someone who privately maintains data in unknown formats, who would you choose as a collaborator?

If one is not chosen as a collaborator, with authorship, what other rewards are available? Here the existing system still falls short. Citation of data sources is not always easy, or enforced, or effective. You receive little credit if all an author can only say Smith (unpublished data). You may not be individually credited if your data is part of a large database that gets cited. A mention in acknowledgements is currently only of minor benefit.

Yet the publish-or-perish reward system is already changing in ways that favor data sharing. Currently, authorship is considered paramount, particularly in journals with high impact factors based on overall journal citation rates. Recently, a measure of individual researcher impact has been proposed (Hirsch 2005). This H index takes into account the number of citations the author's papers have received, data now publicly available at http://scholar.google.com. While how best to compute such an index is still controversial and beyond the scope of this essay, we now have the means to objectively and easily evaluate individual impact.

Recommendations: Journals should make it easier to cite data by allowing extended online citations; journals must enforce rich data citation; and employers and funding agencies should use such citations in evaluating researcher performance. Tracking data citations will allow the effectiveness of data archives and methods to be judged.

In a self-fulfilling prophecy, ecology journals have notoriously low impact factors, perhaps in part because emphasis has been placed on new data collection rather than integration of old data into more modern analyses. Journals that associate papers with well-annotated data can not only increase the impact of individual papers and researchers but can lift their own impact as well.

The future

Years ago I planned but did not publish an essay titled “Ecologists as content providers,” urging ecologists to consider more direct contributions to the World Wide Web. Much has changed in the online landscape so that now ecologists can be citizens in online knowledge-building communities. The rise of open access journals, open source software, and collaborative content-building has challenged old models of intellectual property and the best ways to foster creativity, progress, and quality (Vaidhyanathan 2003, Weber 2004).

In an open source approach to science (e.g. Maurer 2003), exchanges of data will be the rule. If someone finds errors in a shared dataset, as all of us who work with such data have, he can offer a patch to the community (to borrow phraseology from software development). I may transform the data of others into a new format and save others the trouble, as we have done at http://www.spire.umbc.edu. As in complex software projects, scientific communities can mobilize a coordinated open source project towards a shared goal. We are creating one for lepidipteran systematists at http://www.leptree.net.

Not everyone must take an open source approach to data sharing. In the world of software, both private and open source models appear to be sustainable and many of us take advantage of both in our personal and professional lives. There is growing interest in using the Semantic Web as a framework for exchange of data. Though it needs more study, researchers believe it holds promise both for intelligent integration and for addressing complex policy issues of data access and quality (Berners-Lee et al. 2006). The road to the “web of data” described by the W3C (http://www.w3.org/2001/sw/) is likely to be a long and interesting one.

Acknowledgements (if allowed): The author thanks Tim Finin and Charlie Mitter for suggestions. She is supported by NSF ITR 0326460 and NSF AToL 0531769.

Berners-Lee T, Hall W, Hendler J, Shadbolt N, Weitzner DJ. 2006. Creating a science of the web. Science 313(5788): 769-771.

Blackburn TM, Gaston KJ. 1998. Some methodological issues in macroecology American Naturalist 151 (1): 68-83 JAN 1998

Cassey P, Blackburn TM. 2006. Reproducibility and repeatability in ecology. BioScience 56(12):958-959.

Hirsch, JM. 2005. An index to quantify an individual's scientific research output. Proc. Nat. Acad. Sci. 102(46):16569-16572.

Jones MB, Shildhauer MP, Reichman OJ, Bowers S. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Ann. Rev. Ecol. Evol. Syst. 37:519-544.

Maurer SM. 2003. New Institutions for doing science: from databases to open source biology. European Policy for Intellectual Property Conference on Copyright and database protection, patents and research tools, and other challenges to the intellectual property system (http://www.merit.unimaas.nl/epip/papers/maurer_paper.pdf)

National Research Council. 2003. Sharing Publication-related Data and Materials: Responsibilities of Authorship in the Life Sciences. New York: The National Academies Press.

Palmer MA. et al. 2004. Ecological science and sustainability for a crowded planet, Ecological Society of America (http://www.esa.org/ecovisions).

Parr CS, Cummings M. 2005. Data sharing in ecology and evolution. Trends Ecol. Evol. 20(7):362-363.

Vaidhyanathan S. 2003 Copyrights and Copywrongs: the Rise of Intellectual Property and How it Threatens Creativity. New York: NYU Press.

Weber S. 2004. The Success of Open Source. Cambridge, MA: Harvard University Press.