Open Sourcing Ecological Data

CYNTHIA SIMS PARR

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. Cassey and Blackburn 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 reuse

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 they can be reanalyzed. 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, the potential value of data to the community can be a sufficient reason to require their publication. Ecologists should not be subject to a standard lower than that for 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.

I recommend dialogue within the scientific community to help journals and reviewers determine when reproducibility and reusability are most desirable. Then data-sharing requirements can be consistently and fairly applied. The National Research Council 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 manuscripts to them.

Journals as data custodians

Cassey and Blackburn (2006) express concern that journals are not ready to be custodians of original data. Larger publishers (e.g., *Ecology, Science, Nature*) already archive data and supplemental material, largely in flat formats; such data may not be the most easily found or used, but there is significant value in their 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 that are already working with the scientific community to develop standards and protocols (Parr and Cummings 2005, Jones et al. 2006). The best way to speed progress on these data repositories 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, as 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.

My suggestion is that each journal create a consistent policy about where, how, and when the data associated with its publications must be archived. Until clear community standards emerge, journals should provide a wide range of alternatives 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 that demand specialized skills and training. These skills are necessary to effectively integrate and reuse the data, a process fraught with pitfalls (Jones et al. 2006). If others would put your data set to the same use you plan for it, wouldn't you, as the origina-

Cynthia Sims Parr (e-mail: csparr@umd.edu) is an adjunct professor in the Behavior, Ecology, Evolution and Systematics program, University of Maryland, College Park, MD 20742; a research associate at the University of Maryland Institute for Advanced Computer Studies in College Park; and an assistant research professor in the Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore, MD 21250.

tor of the data, 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 that you can see are likely to be useful, and someone who privately maintains data in unknown formats, whom 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, enforced, or effective. You receive little credit if an author can say only "Smith (unpublished data)." You may not be individually credited if your data are part of a large database that gets cited. A mention in the acknowledgments 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. But a measure of individual researcher impact has recently been proposed (Hirsch 2005). This "H index" takes into account the number of citations the author's papers have received, data that are now publicly available at Google Scholar. Although the question of 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.

I believe that journals should make it easier to cite data by allowing extended online citations; that journals must enforce rich data citation; and that employers and funding agencies should use such citations in evaluating researcher performance. Tracking data citations will also allow us to judge the effectiveness of alternative data archives and methods.

In a self-fulfilling prophecy, ecology journals have notoriously low impact factors, perhaps in part because more emphasis has been placed on new data collection than on the 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 raise 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 notions about the best ways to foster creativity, progress, and quality.

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 data set, as all of us who work with such data have, he or she 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 my colleagues and I on the Spire research project have done. As in complex software projects, scientific communities can mobilize a coordinated open-source project toward a shared goal. (We are creating one for lepidopteran systematists at *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. The road to the "web of data" described by the World Wide Web Consortium is likely to be a long and interesting one.

Acknowledgments

The author thanks Tim Finin and Charlie Mitter for suggestions. C. S. P. is supported by National Science Foundation grants NSF ITR 0326460 and NSF AToL 0531769.

References cited

Cassey P, Blackburn TM. 2006. Reproducibility and repeatability in ecology. BioScience 56: 958–959.

Hirsch JM. 2005. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences 102: 16569–16572.

Jones MB, Shildhauer MP, Reichman OJ, Bowers S. 2006. The new bioinformatics: Integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics 37: 519–544.

Maurer SM. 2003. New institutions for doing science: From databases to open source biology. A paper presented to the European Policy for Intellectual Property Conference on copyright and database protection, patents and research tools, and other challenges to the intellectual property system. (15 February 2007; *www.merit.unimaas.nl/epip/papers/maurer_paper.pdf*)

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

Palmer MA, et al. 2004. Ecological Science and Sustainability for a Crowded Planet: 21st Century Vision and Action Plan for the Ecological Society of America. (15 February 2007; *www.esa.org/ecovisions*)

Parr CS, Cummings M. 2005. Data sharing in ecology and evolution. Trends in Ecology and Evolution 20: 362–363.

doi:10.1641/B570402
Include this information when citing this material.