vol. 175, no. 2 the american naturalist february 2010

Editorial

Data Archiving

Michael C. Whitlock,[1] Mark A. McPeek,[2] Mark D. Rausher,[3] Loren Rieseberg,[4] and Allen J. Moore[5]

1. Department of Zoology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada (former Editor-in-Chief, *The American Naturalist*); 2. Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire 03755 (Editor-in-Chief, *The American Naturalist*); 3. Department of Biology, Duke University, Durham, North Carolina 27708 (Editor-in-Chief, *Evolution*); 4. Department of Botany, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada (Chief Editor, *Molecular Ecology*); 5. Centre for Ecology and Conservation, School of Biosciences, University of Exeter, Cornwall Campus, Penryn TR10 9EZ, United Kingdom (Editor-in-Chief, *Journal of Evolutionary Biology*)

Science depends on good data. Data are central to our understanding of the natural world, yet most data in ecology and evolution are lost to science—except perhaps in summary form—very quickly after they are collected. Once the results of a study are published, the data on which those results are based are often stored unreliably, subject to loss by hard drive failure and (even more likely) by the researcher forgetting the specific details required to use the data (Michener et al. 1997). Moreover, most data are never available to the broader community, even after publication of the results; in most cases this unavailability becomes permanent following the eventual death of the researchers involved. In ecology and evolutionary biology, we are losing nearly all of this important legacy.

Yet these data, even after the main results for which they were collected are published, are invaluable to science, for meta-analysis, new uses, and quality control. With the increasing use of meta-analysis to summarize multiple studies, it has become clear that necessary summary statistics are often not published. In many cases, a study can be used only if the original data are available to the meta-analysts. Furthermore, data often can be used in ways beyond the questions that sparked their collection; for example, many studies contain information that can serve later as a baseline for detecting population trends, even decades later. The availability of data for published studies also allows error checking, making science more open and letting us more rapidly reach accurate conclusions. Finally, papers that have had data archived are more useful to—and more cited by—other scientists. One study found that papers that archived their data were cited 69% more often than papers that did not (Piwowar et al. 2007).

Data that are properly archived are saved for posterity, and archives also function to preserve data in a usable form for the original authors. Moreover, if data sets are put into a readily interpretable format while the methods and structure of the data are foremost in the scientists' minds, those data can be used later more easily by those scientists and others.

The example of GenBank shows the value of the availability of data for all of these reasons. The modern synthetic use of DNA sequence data would not be possible without the near-universal use of GenBank as a public archive. Moreover, GenBank would not be nearly as complete as it is without the communal decision to archive all DNA sequence data, a decision initially introduced by journals.

For these reasons and perhaps others, a survey has shown that over 95% of scientists in evolution and ecology think that data should be publicly archived (S. Carrier, J. Greenberg, H. Lapp, R. Scherle, A. Thompson, T. Vision, and H. White, unpublished manuscript).

To promote the preservation and fuller use of data, *The American Naturalist*, *Evolution*, the *Journal of Evolutionary Biology*, *Molecular Ecology*, *Heredity*, and other key journals in evolution and ecology will soon introduce a new data-archiving policy. The policy has been enacted by the Executive Councils of the societies owning or sponsoring the journals. For example, the policy of *The American Naturalist* will state:

> This journal requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as GenBank, TreeBASE, Dryad, or the Knowledge Network for Biocomplexity. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Authors may elect to have the data publicly available at time of publication, or, if the technology of the archive allows, may opt to embargo access to the data for a period up to a year after publication. Exceptions may be granted at the discretion

Am. Nat. 2010. Vol. 175, pp. 145–146.
0003-0147/2010/17502-51762$15.00.
DOI: 10.1086/650340

of the editor, especially for sensitive information such as human subject data or the location of endangered species.

This policy will be introduced approximately a year from now, after a period when authors are encouraged to voluntarily place their data in a public archive. Data that have an established standard repository, such as DNA sequences, should continue to be archived in the appropriate repository, such as GenBank. For more idiosyncratic data, the data can be placed in a more flexible digital data library such as the National Science Foundation–sponsored Dryad archive at http://datadryad.org.

When the policy is fully in place, authors will archive the data required to support the conclusions in their published paper, along with sufficient details so that a third party can reasonably interpret those data correctly. In most cases this will require a short additional text document, with details specifying the meaning of each column in the data set. The preparation of such shareable data sets will be easiest if these files are prepared as part of the data analysis phase of the preparation of the paper, rather than after acceptance of a manuscript.

The data should be saved usually at the individual level, although what is most important is that the data be saved in a way that makes the most sense for later usability. Summary statistics (like means and standard deviations) are not sufficient, because they wouldn't provide enough information for later analysis. At the same time, data in their rawest form, such as videotapes, field notebooks, or sequencing trace files, are not required. For example, if a study used videotape to determine the time required for animals to make a choice in a T-maze, the data should be recorded with the time for each subject.

The data-archiving policy is designed to address several concerns that some researchers may have about data sharing. To protect the ability of individual researchers to use the data that they have collected, the policy allows an embargo period after publication. While the data will be entered into an archive at the time of publication, the data may be restricted from public view for up to a year. This allows the original researcher time to publish other papers based on the data set. The policy also allows longer embargo periods at the discretion of the editor in exceptional cases. In addition, the requirement is only for data that have already been used in the publication in question; other data from the same research project that have not yet been described in a publication need not be archived. Finally, data that are particularly sensitive, such as location information for endangered species subject to poaching, should not be archived in a publicly accessible format. Human subject data should be anonymized (see the recommendations of the National Human Subjects Protection Advisory Committee 2002).

Throughout the history of ecology and evolution, enormous quantities of valuable data have been lost to future science, for a variety of technical and cultural reasons. There are no longer any meaningful technical barriers to long-term storage, and just as in the case of DNA sequence data, it is time for the culture of our shared use of data to evolve. With this general data policy, we think that science will reap great benefits for generations to come.

Literature Cited

Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford. 1997. Nongeospatial metadata for the ecological sciences. Ecological Applications 7:330–342.

National Human Subjects Protection Advisory Committee. 2002. Recommendations on public use data files. http://www.aera.net/humansubjects/NHRPAC_Final_PUDF.pdf.

Piwowar, H. A., R. S. Day, and D. B. Fridsma. 2007. Sharing detailed research data is associated with increased citation rate. PLoS ONE 2(3):e308, doi:10.1371/journal.pone.0000308.