Data Sharing in Ecology and Evolution: Why Not?

Cynthia S. Parr[1] and Michael P. Cummings[2]

[1]Institute for Advanced Computer Studies, [2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, 20742 USA

Corresponding author: Parr, C.S. (csparr@umd.edu)

The rapid advancement of fields such as molecular biology, genomics, and molecular evolution is in large part due to pervasive data sharing. New discoveries are made through application of bioinformatics to widely available datasets. Ecology, evolution, conservation biology, animal behavior and related fields have not enjoyed similar growth. Data sharing can transform these fields, as it has others [1,2], but first individual scientists must recognize the benefits and see their way past perceived barriers.

Several compelling ethical reasons for data sharing apply to ecology and evolution. As in other fields, sharing data that supports publications, in useful formats and in community-accepted archives, facilitates the scientific ideals of replication, building on prior work, and synthesis [1].   Also, most research in ecology and evolution is publicly funded, so one might argue that the data belong to the public. Sharing data provides additional return on that investment. For example, the *Iris* flower measurement data of Anderson [3] were used soon after publication by Fisher [4] to illustrate discriminant functions, and decades later are probably the most-used data in machine learning research. Given the already scarce funding in ecology and evolution, if data to answer a new question already exist, why spend time and money to collect it again? The larger implications of data sharing are also important. Can researchers morally justify keeping data private if these data may speed solutions to environmental and conservation challenges? Participants in the new Conservation Commons Initiative (http://conservationcommons.org) think not.

Ecologists and evolutionary biologists are getting better at data sharing (Table 1), although we certainly have not yet achieved Ellison's [5] "Tapestry of Nature."   Relatively few researchers are participating, and it remains difficult to explore and use these repositories and registries.

Why is data sharing not yet common practice? A recent NCEAS survey is exploring attitudes in detail (S. Findlay, *pers. comm.*) but two obvious reasons cited in [6] are that researchers desire to use their data for subsequent work without competition, and they believe there are logistical barriers to data sharing.

Withholding data for possible future gain is shortsighted, because the academic reward system favors data sharing. The value of data increases in proportion to its use by others, with direct consequences in perceptions of research importance and in objective measures (e.g. citation rate). These perceptions and measures are used formally and informally as criteria for publication, grant funding, and career advancement.

Logistical barriers to data sharing are illusory. Convenient means to share data already exist even for researchers not associated with large scale efforts such as international Long-term Ecological Research projects. One may submit supplementary files to journals, post data on institutional web sites such as the Digital Repository at the University of Maryland (http://drum.umd.edu), or simply post files on project web sites. Infrastructure and tools [7,8] are being developed that support data sharing in and use of formal ecological repositories and registries (see Table I). Ecological societies are working to achieve consensus on institutional goals and policies related to data sharing [6,9]. New methods of discovery and automated data integration (e.g. [10]; the ORIEL project, http://www.oriel.org/), can take advantage of ontologies for animal behavior (http://ethodata.nsdl.cornell.edu), animal life history (http://animaldiversity.ummz.umich.edu/site/about/technology/), and ecology (http://wow.sfsu.edu/ontology/rich/). Active data sharing itself fosters increased standardization, as the best-annotated or -collected data are more likely to be re-used and cited.

Ecology and evolution should be part of the larger synthetic, multidisciplinary movement (e.g., how do ecological processes affect the epidemiology, etiology, and vulnerabilities of emerging diseases? [11]). In the United States, researchers at NCEAS and, soon, NESCENT and the National Ecological Observatory Network are forging ahead with exciting research that relies on shared data. Data shared as benchmark datasets (e.g. [12]) can kick-start innovation by providing well-defined challenges to computer scientists and informatics experts. The resulting technology can speed progress by ecologists and evolutionary biologists.

With substantial benefits for individuals, scientific communities, and society as a whole, the time for data sharing has come. It is up to us individually to take advantage of the many opportunities to share data, to make use of that data, and to support the development of improved tools and techniques for working with that data. Why not?

Reference List

1 National Research Council (2003) Sharing publication-related data and

materials: responsibilities of authorship in the life sciences, Washington, DC, The National Academies Press

2 Insel, T.R. *et al.* (2003) Neuroscience networks: Data-sharing in an information age. *PLoS Biology* 1, 9-11

3 Anderson, E. (1935) The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society* 59, 2-5

4 Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188

5 Ellison, A.M. (1998) Data: the tapestry of nature. *EcoEssay Series Number 2*, National Center for Ecological Analysis and Synthesis. Santa Barbara, CA. (http://www.nceas.ucsb.edu/fmt/doc?/nceas-web/resources/ecoessay/ellison/)

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

7 Michener, W.K. (2003) Building SEEK: the Science Environment for Ecological Knowledge. *DataBits: An electronic newsletter for Information Managers* 3 (http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/03spring/)

8 Cotter G. *et al.* (2004) Integrated science for environmental decisionmaking: the challenge for biodiversity and ecosystem informatics. *Data Science Journal* 3, 38-59

9 Silver, S. (2004) Editorial: Publishing for the digital age. *Frontiers in Ecology and the Environment* 10, 507

10 Caragea, D. *et al.* (2004) Learning classifiers from semantically heterogeneous data. Proceedings of the Third International Conference on Ontologies, DataBases and Applications of Semantics for Large Scale Information Systems (ODBASE’04),

11 American Institute of Biological Sciences (2004) Ecology and evolution of infectious diseases. Report from a NEON Science Workshop. AIBS, Washington, DC. (http://www.neoninc.org/documents/neon-disease-report.pdf)

12 Plaisant, C. (2004) The challenge of information visualization evaluation. Procdeedings of Advanced Visual Interfaces 2004, Pp. 320-327, ACM Press

Table 1. Examples of ecological and evolutionary data registries (providing access to metadata and pointers to data stored elsewhere), institutional repositories (archived datasets), and topical repositories (specific kinds of archived datasets in standardized file formats). This list includes only sources with online access to machine-readable data and metadata; datasets are counted or self-reported as of 23 February 2005. Datasets in repositories may also be represented in registries.

Registries	**Datasets**	**URL**
Global Biodiversity Information Facility portal	343*	http://www.gbif.org
National Biological Information Infrastructure	17000	http://www.nbii.gov
Knowledge Network for Biocomplexity	1500**	http://knb.ecoinformatics.org/index.jsp
Institutional or journal repositories		
NCEAS	72	http://knb.ecoinformatics.org/knb/style/skins/nceas/
Ecological Archives data papers (ESA journals)	6	http://www.esapubs.org/archive/archive_D.htm
Topical repositories		
Interaction Web Database	74 webs	http://www.nceas.ucsb.edu/interactionweb
TreeBase	2869 phylogenetic trees	http://www.treebase.org/treebase/
Global population dynamics database	5000 time series	NERC Centre for Population Biology http://cpbnts1.bio.ic.ac.uk/gpdd/
VegBank	19000 plots	http://vegbank.org/vegbank/index.jsp

*data sets are typically museum collections -- total records over 45 million

**Includes Long-term Studies Section Data Registry (of ESA), e.g. 567 datasets from LTER, 434 from Univ. of California Natural Reserve System, 193 from Organization of Biological Field Stations