Building Support for a Discipline-Based Data Repository

Ryan Scherle[1], Sarah Carrier[2], Jane Greenberg[2], Hilmar Lapp[1], Abbey Thompson[2], Todd Vision[1,3], Hollie White[2]
[1]National Evolutionary Synthesis Center (NESCent), Durham, NC
[2]School of Information and Library Science, University of North Carolina, Chapel Hill, NC
[3]Department of Biology, University of North Carolina, Chapel Hill, NC

The field of evolutionary biology is suffering from a crisis of data attrition. Although specialized repositories (such as GenBank) exist for some of the most commonly seen data types, it is rare that every dataset associated with a published paper has a suitable permanent home. Furthermore, while many evolutionary biology journals have policies that encourage authors to share data, evolutionary biology is typical of many "small science" disciplines in that there are only piecemeal standards, and little infrastructure, that enable authors to do so. At the behest of major journals and societies in evolutionary biology, we have begun development of a repository called Dryad for the preservation, discovery and sharing of data underlying published works in the field.

1 Existing Repositories

Evolutionary biology is a highly interdisciplinary field, combining research from areas such as ecology, developmental biology, genetics, molecular biology, paleontology, and systematics. Due to this diversity, data comes in a wide variety of formats, ranging from simple ASCII text (e.g., gene sequences) to tabular data (e.g., spreadsheets), to images, to complex data such as videos and simulations.

Evolutionary biology is not without its own repositories. GenBank (together with its European and Japanese counterparts EMBL and DDBJ, respectively) is the preeminent database for genetic sequences [1]. GenBank has been highly successful, and most journals (as well as federal funding agencies) require authors to submit sequences to GenBank. Another well-established repository required by some journals is TreeBASE, which specializes on phylogenetic trees, representing the evolutionary relationships between species [2]. By specializing on particular types of data, these repositories can have highly structured data, rich metadata, and analytical capabilities uniquely tailored to their contents. However, these repositories are largely insular, and do not provide standardized mechanisms for sharing data.

Organism-oriented repositories, such as FishBase [3], collect data on targeted sets of species. These repositories accept a wider variety of data formats and make sophisticated links between them, but their use is limited to a few small research communities.

Many journals in the field allow authors to upload "supplemental material" associated with a publication. The supplemental material repositories allow nearly any type of file. However, these systems have many deficiencies. They do not provide standard repository features such as rich metadata or persistent identifiers. They often place limits on the number and size of files that may be submitted. In many cases, authors simply use the supplementary materials repository to provide an extra figure or table, rather than detailed data underlying the figures and tables already in the article.

2 Attitudes Towards Data Sharing

In 2006, The StORe project conducted a survey of repository use across a broad range of disciplines [4]. They found that repository use is particularly prevalent in the biosciences (although the study did not differentiate evolutionary biology from other subdisciplines), and use of GenBank is well established. However, much data is still being lost.

Studies have shown that informal procedures for data sharing (not involving repositories) are problematic. In a survey of 1240 geneticists [5], 47% had been denied at least one request for data or materials in the preceding three years, and 28% reported that they had been unable to confirm published research because of data withholding. The primary reasons given for withholding data were that too much effort was required to collect the requested data (80% of respondents), or authors were protecting their ability (57%) or the ability of a colleague (64%) to publish further results.

As an additional challenge, even when portions of data are made available, there is no guarantee that sufficient information is available to permit replication of scientific results. For example, many entries in TreeBASE lack critical details about the gene sequences and sequence alignments used to generate the tree.

To determine the extent and kinds of data at risk of loss in the field of evolutionary biology, we examined 27 randomly selected articles from five major journals. While data was provided in the journal's supplemental data repository for 41% of the articles, it was typically composed of similar types of data objects as in the paper itself. Raw data tables were rarely provided. For example, the majority of studies (67%) that made use of sequence alignments (a common datatype) did not make the alignments available. The vast majority of articles in our sample (78%) were based at least in part on the analysis of datasets not deposited in any repository.

Authors often take the attitude that "no one else will be interested in my data, so there is no reason to share it", yet these same authors frequently make use of data published by others. 48% of the articles in our sample based conclusions on data from previously published studies.

3 Working With Journals

From the inception of the Dryad project, journal editors in the field have been kept in close consultation. The editors are key stakeholders. They are active scientists themselves and so are sensitive to their authors' needs and concerns. Journal editors can summarize the requirements of authors. In addition, journal cooperation is critical if we wish to collect data at the time articles are accepted for publication, and if we expect journals to eventually require submission of data for publication.

In December 2006, a small group of journal editors and scientific society representatives participated in a workshop at NESCent to articulate initial requirements for Dryad and formulate a strategy for developing and promoting the repository. A larger workshop was held in May 2007, including a mixture of journal and society representatives and information science professionals. Results from these workshops included:

- Immediate data collection to stem the loss of data is top priority. Detailed metadata curation and analysis tools can be introduced later.
- To ensure credibility, there must be a focus on data associated with published articles.
- A cumbersome data submission process presents a serious barrier to submission. Metadata collection must be automated as much as possible.
- The repository must not dictate any specific data format.
- Journals must be responsible for encouraging deposition and verifying the quality of deposits.
- Open policies for data access (CreativeCommons or ScienceCommons) should be put in place.
- It is critical to develop good practices for data citation and provide automated support for those practices.
- Repository contents must be available via a variety of open standards, including OAI-PMH and SRU.

In addition to achieving these points of consensus, the journal representatives pledged to work towards adoption of a joint data sharing policy by their respective societies or journals. The draft of that policy reads as follows:

"<<Journal title>> requires, as a condition for publication, that data used in the paper should be archived in an appropriate public archive, such as GenBank, TreeBASE, or Dryad. The data should be given with sufficient details that, together with the contents of the paper, allows each result in the published paper to be re-created. Authors may elect to have the data publicly available at time of publication, or, if 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 of the editor, especially for sensitive information such as the location of endangered species."

Current plans call for simultaneous adoption of this policy by the major journals in the field, which will ensure that no one journal suffers from decreased submissions due to its adoption.

In January of 2008, we began a round of presentations to the executive councils of the societies and editorial boards of our target journals to ensure community engagement. The target journals are being asked to appoint a representative to a Dryad Management Board, which will be responsible for major policy and strategy decisions, including guidelines for use and citation of data in the repository, elective embargo periods for authors, and the financial model for long-term sustainability. In addition, journals are currently being asked to encourage voluntary deposition of data for authors of accepted publications. To date, there has been strong support for joint adoption of the deposition policy when Dryad is fully functional.

4 Working With Researchers

As we work with the journals and societies, we are also studying the community of scientists that Dryad is being designed to serve. As a first step, we are conducting a survey to gain a broad understanding of community practices and preconceptions regarding repository use. The survey is conducted via the web-based SurveyMonkey platform, allowing us to easily obtain participation and tabulate results.

To obtain a more detailed view of how individual researchers approach the problem of data archiving, we are conducting a use-case study. In this study, individual researchers are being interviewed to collect detailed descriptions of their approaches to data management, archiving, and distribution, as well as their use of data published by other researchers.

5 References

[1] D.A. Benson, et al., GenBank. *Nucleic Acids Research,* January 2007, 35(Database issue):D21-5.

[2] W.H. Piel, et al., TreeBASE: a database of phylogenetic information, *Second International Workshop of Species*, Tsukuba, Japan, 2000.

[3] R. Froese and D. Pauly, Editors, *FishBase 2000: concepts, design and data sources*. ICLARM, Los Baños, Laguna, Philippines. 2000.

[4] G. Pryor, Attitudes and aspirations in a diverse world: the Project StORe perspective on scientific repositories, *Second International Digital Curation Conference*, Glasgow, November 2006.

[5] E.G. Campbell, et al., Data withholding in academic genetics: evidence from a national survey, *Jama*, 2002, 287(4): p. 473-80.