# *"Data Preservation, Sharing, and Discovery: Challenges for Small Science in the Digital Era"*

## A report of a workshop held May 16 -17th 2007 at NESCent, Durham, NC

**Sponsors:** The National Evolutionary Synthesis Center (NESCent) and the University of North Carolina at Chapel Hill, School of Information and Library Science, Metadata Research Center (<MRC>)

**Organizers:** Jane Greenberg (MRC), Hilmar Lapp (NESCent) and Todd Vision (NESCent).

**Participants:** Participants included project participants from the NESCent-MRC digital data sharing initiative, representatives from major evolutionary biology journals and societies, and experts from the information, library and computer science communities.

Representatives and trainees from UNC/SILS/MRC

- MRC students and postdocs: Amy Bouck, Sarah Carrier, Jed Dube
- MRC scribes: Lori Eakin, Carolyn Hank, Laura Sheble

Representatives from NESCent, journals and scientific societies

- Bradley Anholt, University of Victoria/Journal of Evolution Biology
- Ahrash Bissell, Duke University/OpenContext
- Harold Heatwole, NCSU/Editor of Integrative and Comparative Biology
- Joel Kingsolver, UNC Chapel Hill/NESCent/American Society of Naturalists
- Mark Rausher Duke University/Editor of Evolution
- Loren Rieseberg, University British Columbia/Editor of Molecular Ecology (participating by conference call)
- Kathleen Smith, Duke University/NESCent
- Marcy Uyenoyama, Duke University/Editor of Molecular Biology and Evolution
- Derek Wildman, Wayne State University, Assistant Editor, Molecular Phylogenetics and Evolution

Representatives from Information, Library and Computer Science

- Leesa Brieger, Renaissance Computing Institute
- Brad Hemminger, SILS/UNC Chapel Hill
- Paul Jones, iBiblio and SILS/UNC Chapel Hill
- John Madden, Duke University
- Kristin Martin, UNC Chapel Hill Libraries
- Steve Morris, NCSU Digital Libraries Initiative
- Michael Nelson, Old Dominion University and the Open Archives Initiative
- Oya Rieger, Cornell University Libraries
- David Romito, UNC Chapel Hill Libraries
- Ryan Scherle, Indiana University/NESCent

- Gail Steinhard, Cornell University Libraries
- Helen Tibbo, SILS/UNC Chapel Hill
- Larry Tindell, private contractor/NIEHS
- Stuart Weibel, OCLC Online Computer Library Center/Dublin Core Metadata Initiative

**Problem:**

In "small science" disciplines, such as evolutionary biology, data are typically collected by individual investigators and are highly heterogeneous in content and structure. The nature of these data, and current information infrastructure limitations, pose unique challenges to preservation, discovery, sharing and integration.

**Goals:**

The overarching goals of this workshop were to: 1) identify the way forward for long-term preservation and sharing of digital datasets underlying published works in evolutionary biology, 2) inform the design of Dryad, the NESCent-MRC data repository project, and 3) discuss solutions portable to other small science data repository initiatives.

**Workshop format:**

The workshop ran two days and included themed issue-driven breakout groups, breakout group reporting, and an all-participant discussion each day. The latter part of the second day included a special meeting with journal representatives.

**Summary of Activities and Discussions:**

**Day 1:**

The meeting opened with a welcome to NESCent and introductions. This was followed by a set of brief presentations to provide background information useful for the breakout discussions:

- Todd Vision: Introduction to DRIADE
- Ahrash Bissell: OpenContext and the Alexandria Archive Institute
- Leesa Brieger: SRB & iRODS
- Paul Jones: iBiblio.org
- Michael Nelson: OAI and OAI standards
- Oya Rieger: Data Curation and eScholarship
- Gail Steinhart: Integrating discipline-specific metadata standards for digital curation
- Stuart Weibel: OCLC, Identifiers, Web 2.0
- Helen Tibbo: Institutional Repositories
- Jane Greenberg: Workshop objectives

Four concurrent breakout groups, addressing the following topics and questions, followed the opening session:

1. Adoption and sustainability. How can a repository achieve long-term financial sustainability? What must be done to promote adoption by scientists?

2. Intellectual property. How to balance intellectual property concerns with the goals of data sharing?
3. Distribution and replication. What are the possibilities and liabilities of distributed data management and how can a repository interface with specialized repositories?
4. Lifecycle management. What are the challenges and issues related to a heterogeneous and changing landscape of data formats and technologies?

The breakout group session was followed by breakout group reports, and an all group discussion of key theses and challenges requiring further discussion on day 2.

**Day 2**

Day 2 opened with an all group discussion following on key emerging issues from day 1, and a restructuring of breakout groups themes and questions. Topics and questions addressed by breakout groups included the following:

1. Promoting participation. What would make deposition attractive for depositors from a small science community, and what functionality would maximize the use of the deposited data?

2. Lessons learned from other projects. Who can we learn from, and what lessons do we take from them, in particular with respect to sustainable funding models, approaches to data acquisition and incentivization of users?

3. Architecture. What are the best practices in technical management of data, metadata, and identifiers over the metadata lifecycle? How can we nurture the bottom-up growth of standards for data and metadata formats, and for inter-repository interoperability?

The workshop closed with a group discussion on emerging technologies (e.g. semantic web/Web 2.0) and their potential impact for digital data sharing in small science disciplines. Stuart Weibel of OCLC initiated the session with a brief presentation on Web 2.0 technologies.

All discussions (breakout groups and all-participant discussions) were recorded and written summary notes were produced by SILS/MRC students. Discussion notes and summaries are available from:
http://driade.nescent.org/Public:DRIADE_Workshop_May_2007

After general participants departed, a stakeholder discussion was held that included DRIADE personnel (Bouck, Greenberg, Lapp, Smith, Vision) along with editors and society representatives (Anholt, Kingsolver, Uyenoyama, Wildman, Rieseberg).

**Conclusions:**

The recommendations to emerge from the discussions can be broadly classified into the areas of Sustainability, Policy and Responsibilities, Adoption and Technical Infrastructure.

*Sustainability*: A disciplinary repository focused on published data needs to include a critical mass of highly respected journals on order to achieve community support and to be financially sustainable. The business model needs to ensure long-term viability, and user trust, and the management structure needs to ensure community oversight and ownership. A variety of financial models are possible, but some participants voiced the concern that charging for data access could pose a risk to adoption. The effort should also engage students to use shared data and tools and acclimatize them to the culture and value of data sharing.

*Policy and Responsibilities:* Journals need to take responsibility for author compliance with data sharing policies, including setting clear policies as to what data objects need to be deposited, and enforcing data deposition. Repositories need to have human curators to validate submissions, perform metadata quality control and maintenance, and to help users with problems. The policies for data use and citation need to be clearly articulated. There was strong support for the explicit, easily communicated, and inherently open, approach of Creative Commons and Science Commons, though considerable debate still exists over the role of legal strictures versus community norms in guiding user behavior. Additionally, a data sharing policy must permit exceptions for sensitive data (whether due to privacy, conservation concerns, etc).

*Adoption*: A repository capable of handling highly heterogeneous submissions needs to exist first before some of the journals will be able to require data sharing from their authors. It is important for community acceptance that there be a way for data to be cited, and for data users to also credit the original creators of the data. It was agreed that allowing for data embargoes of limited time would assist with acceptance. One-stop data deposition would be a major incentive to researchers, and stakeholders agreed that the first repositories to consider for inclusion would be Genbank, Treebase, and possibly repositories for microarray data. The deposition process itself needs to be extremely simple. Providing tools and services for scholarly communication, analysis, such as citation counters, format converters, etc, have been found to promote repository adoption. A successful system will incorporiate the researcher's work process. It was pointed out repeatedly that changing a scientific culture is very difficult. It was noted that there is essentially no research on the motivations and practices of data sharing and withholding in the field of, evolutionary biology. and that such research would be helpful in setting priorities as well as identifying challenges and opportunities.

*Technical infrastructure:* Globally unique, persistent, and resolvable identifiers are crucially important for data repositories. In the case of Dryad, identifiers would be desirable both for the entire "data package" associated with a single published journal article, and for each of the individual data objects—both as first class objects. Despite agreement on the importance of identifiers, consensus on the appropriate technology was not achieved. Adoption of metadata and vocabulary standards, metadata harvesting standards (such as OAI-PMH), and web service standards (such as SRU) will allow for much greater interoperability with other repositories and digital libraries. Given the overlapping scientific communities, interoperability with existing ecoinformatics tools (e.g. MetaCat) should be a priority. Consideration should be given to long-term preservation technologies and models for repository replication (e.g. LOCKSS). The use of full-text from the article would be valuable for providing descriptive and contextual metadata. It is highly desirable for publishers to permit content extraction, or at least selective extraction of journal article content. The repository should capitalize on its unique collection of data to enable Web 2.0 functionalities and features, such as annotation, recommendations, news feeds (e.g., RSS) syndication of updates, deep semantic linking and interconnectivity. Thought should be given to ways that the system could track and take advantage of patterns of user behavior to guide resource discovery and personalize the user interface, for example by providing tag clouds or click statistics.

**Follow-up and outcomes**

The workshop and associated planning sessions helped to crystallize a Joint Data Archiving Policy, drafted by Michael Whitlock (Editor-in-Chief of The American Naturalist), which 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."

The Joint Data Archiving Policy is, at the time of writing, under review by the boards of the participating scientific societies and journals. NESCent has committed to follow up with these boards, and to present the policy to any additional stakeholders that may be identified, with the aim of reaching consensus on a statement that can broadly guide data submission and use policy for Dryad.

A major grant proposal for a digital data repository was submitted in July 2007 to the National Science Foundation program in Biological Databases and Informatics. The proposal would allow Dryad to implement many of the recommendations from the Small Data Workshop. Letters of support were provided from representatives of the following journals: A*merican Naturalist*, *Integrative and Comparative Biology*, *Journal of Evolutionary Biology*, *Molecular Biology and Evolution*, *Molecular Ecology*, *Molecular Phylogenetics and Evolution*, and *Systematic Biology*. This level of community engagement was made possible, in good measure, by the workshop and surrounding discussions. The proposal, if funded, will support further development, implementation and evaluation of a repository for data supporting published research in evolutionary biology, and also allow additional dialog and knowledge sharing in the form of community engagement workshops and outreach activities.

Summary of proposal submitted to NSF BD&I program in July 2007; PI: K. Smith, co-PIs: K. Antelman, J. Greenberg, W. Michener, W. Piel; Duration: 36 months starting 6/1/2008.

## A Digital Repository for Preservation and Sharing of Data Underlying Published Works in Evolutionary Biology

The field of evolutionary biology is suffering from a crisis of data attrition. Specialized databases exist for some of the most commonly seen data types, but 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, NESCent has begun development of a digital repository, called Dryad, for the preservation, discovery and sharing of data underlying published works in the field. The overall aim in this proposal is to facilitate data sharing upon publication by the evolutionary community through a myriad of technical and organizational improvements to Dryad, particularly with respect to (1) the deposition and access interface, (2) incentives and interoperability, and (3) sustainability. We will also (4) promote the use Dryad within the research community, and as an educational tool to teach future scientists about the value of digital data archives. The Specific Aims (SA) include the following.

1. *Deposition and access interface.* [SA1.1] Data deposition will be coordinated with manuscript submission so that reliable bibliographic metadata is automatically captured by Dryad and the identifier for each data object can be transmitted to the journal. [SA1.2] We will study ways to capture more extensive metadata from authors using automatic metadata generation. [SA1.3] To maintain both data integrity and metadata quality, data curators will validate and, if necessary, edit, submissions. [SA1.4] The retrieval interface will structure and augment queries using the scientific metadata coupled with existing and newly developed vocabularies. [SA1.5] Both the fundamental and novel aspects of Dryad will be extensively evaluated and user-tested.
2. *Incentives and interoperability.* [SA2.1] As proof-of-concept for the idea of "one-stop data deposition", we will implement hand-shaking mechanisms with two specialized databases (Genbank, for sequence data, and TreeBASE, for phylogenetic data) such that data will simultaneously be deposited in Dryad and the specialized database. [SA2.2] Dryad will assign globally unique, stable, and resolvable identifiers for datasets and promote a convention for the data citations. [SA2.3] Federation of searches across Dryad and other digital collections in biology and beyond will be achieved by implementing the OAI-PMH protocol for metadata harvesting in Dryad, TreeBASE and Metacat, the premier metadata registry and data repository for ecology. [SA2.4] On-the-fly queries, and syndication, of repository contents will be achieved by implementing the SRU/W standards for web services within Dryad and MetaCat.
3. *Sustainability.* [SA3.1] Dryad will be overseen by a Governing Board (MB) of stakeholders from evolutionary biology journals and societies, advised by information science experts and representatives from other scientific data sharing initiatives, who will set policy and plan for the financial self-sufficiency of the repository beyond the life of

this project. [SA3.2] We will explore technical advances in the long-term stewardship of digital data collections by implementing the LOCKSS distributed data preservation system, in addition to managing a more standard architecture of redundant production and backup systems within the North Carolina State University Libraries.

4. *Community engagement*. [SA4.1] Datasets of special educational value will be targeted and developed for classroom use through a dedicated education section of the repository. [SA4.2] Dryad tutorials will be presented at major evolutionary biology conferences to promote adoption and increase the extent and quality of the metadata provided by authors. [SA4.3] Annual workshops will be held to support emerging metadata and interoperability standards in the field of evolutionary biology, and plan for future handshaking efforts.

The work proposed here will have a broad and transformative impact by enabling the preservation, discovery, sharing and reuse of data for an entire biological discipline. It represents a unique collaboration among diverse institutions (academic journals and associated scientific societies, a national synthesis center and research network, a major community database) and expert communities (evolutionary biologists, information scientists and research librarians). It will serve as a model for the many other "small science" disciplines facing similar challenges in data preservation and sharing.