

## **The Dryad Digital Repository: Published evolutionary data as part of the greater data ecosystem**

Todd Vision

Department of Biology, University of North Carolina at Chapel Hill, and the U.S. National Evolutionary Synthesis Center

Scientific findings are built on a foundation of data. Even though much of that data is now “born digital”, there are numerous technical and socio-cultural challenges to ensuring that digital data are mobilized in a form that can be, on the one hand, discovered, accessed, and processed by machines and, on the other, understood and effectively reused by humans. Scientific data are massively dispersed; even much of the data underlying results reported in the scientific literature is not available in any public repository. The quality of data management varies widely, with idiosyncratic formats and semantics the norm for individual research groups. Many repositories have little guarantee of long-term data preservation.

Dryad [1] is a digital repository for data reported in the scientific literature. It grew out of a data sharing initiative among a consortium of journals in evolution and ecology. Dryad provides a permanent, stable, curated, and updated repository for digital data that otherwise lack a home. It adheres to Open Data principles by making all data available through a CC0 public domain dedication, and promotes data citations by assigning Digital Object Identifiers to data. Dryad also enhances the metadata provided by authors and journals in order to promote data discovery. Dryad works to make data deposition as technically simple as possible through close integration with the manuscript handling systems used by partner journals.

DataONE (Observation Network for Earth, <https://dataone.org>) is a network of data centers that aims to promote data-intensive knowledge discovery by providing open, persistent, robust, and secure access to well-described and easily discovered data about life on earth and its physical environment. DataONE member nodes manage data covering a wide range of spatial and temporal scales and science domains, including citizen-science projects such as eBird and the National Phenology Network. A small number of coordinating nodes provide a distributed metadata registry and core services to users and to what will eventually be a larger network of member nodes, which are themselves replicated across the network for persistent availability and backup. The first set of member nodes to come online include Dryad, the Knowledge Network for Biocomplexity, the Oak Ridge National Laboratories Distributed Active Archiving Center, and the USGS National Biological Information Infrastructure. Integration of these systems demonstrates the feasibility of a common service interface to data managed using a diversity of platforms (e.g. Mercury, MetaCat, and DSpace) and standards (e.g. Ecological Metadata Language, those of the Federal Geographic Data Committee). Here, I describe how the Dryad repository operates within this larger data ecosystem, and illustrate the value of integrating data from published works in evolutionary biology with the wider network of digital data.

### **References:**

[1] Dryad Digital Data Repository, code available under a New BSD license from <http://code.google.com/p/dryad/>