DRIADE: A Data Repository for Evolutionary Biology

Jed Dube, Sarah Carrier, Jane Greenberg

School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC

Introduction

The field of evolutionary biology draws from ecology, paleontology, population genetics, physiology, systematics, and new biological sub disciplines such as genomics. In fall 2006, NESCent (National Evolutionary Synthesis Center) began developing DRIADE (Digital Repository of Information and Data for Evolution) to address fundamental research challenges in evolutionary biology. This poster highlights activities to define functional requirements and a phased development plan.

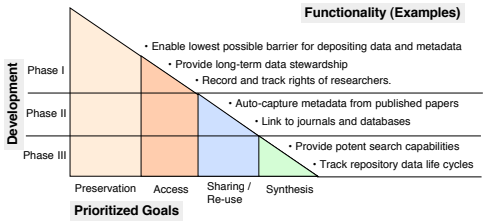
Methods

Approaches used to gather functional requirements:

- Met with key stakeholders (including representatives of evolutionary biology journals) to discuss goals and priorities.
- Conducted a survey/analysis of selected leading digital data and resource repository initiatives, specifically focusing on their applicability to the initial DRIADE goals and priorities.
- Began characterizing data and metadata associated with published articles from selected evolutionary biology journals.
- Employed a multi-method approach to develop a metadata application profile to accompany the initial phase of data collection, preservation, and access.

Prioritized Goals / Phased Development

Following the December 2006 stakeholders meeting, we defined functional requirements and a phased development plan, based on this hierarchy of prioritized goals:



Survey Findings

The following scientific digital data projects and initiatives were considered most useful in helping the DRIADE team understand how to address specific functional priorities for the repository:

Projects: / Goals and priorities:	Global Biodiversity Information Facility (GBIF)	Knowledge Network for Biocomplexity (KNB) / Science Environment for Ecological Knowledge (SEEK)	National Science Digital Library (NSDL)	Interuniversity Consortium for Political and Social Research (ICPSR)	Marine Metadata Initiative (MMI)
Accommodate heterogeneous digital datasets	—	—	—	—	—
Respect intellectual property rights	—	—		—	
Provide tools and incentives to researchers	—	—	—	—	—
Minimize technical expertise and time required by contributors for deposit and access	—	—		—	—
Provide long-term data stewardship	—		—		
Focus on published datasets					

The survey findings enabled definition of specific functional requirements in several key areas; however, we then needed to supplement it with a clearer understanding of published datasets in the evolutionary biology domain.

Characterization of Published Data

We conducted a preliminary analysis of sample data and metadata associated with journal articles, because DRIADE wants to collect and preserve heterogeneous data associated with published papers. We looked at the following:

- Types of data (e.g. protein sequence, map, illustration)
- Forms of data (e.g. tabular, graphic, textual)
- Formats of data (e.g. NEXUS, FASTA, Excel)
- Locations of data (e.g. journal, database)
- Keywords used by authors

The findings further informed the functional requirements, our research plan, and development of the metadata application profile.

Metadata Application Profile

We designed an application profile to support the first phase of DRIADE's development, based on a requirements assessment, content analysis, and crosswalk analysis.[1]

Namespace:Name/Label	Obligation	Generation Method	Occurrences R=Repeatable NR=Non-repeatable
Module 1: Bibliographic Citation			
dcterms:bibliographic Citation/Citation Information	Required	Automatic	R
dc:identifier/Digital Object Identifier	Required	Automatic	NR
Module 2: Data Object			
dc:creator/Name	Required	Semi-Automatic	R
dc:title/Data Set Title	Optional	Manual	NR
dc:identifier/Data Set Identifier	Required	Automatic	NR
PREMIS:fixity/(hidden)	Required	Automatic	NR
dc:relation/DOI of Published Article	Optional	Semi-Automatic or Automatic	R
DDI:<depositr>/Depositor	Required	Manual, then Automatic after profile creation	NR
DDI:<contact>/Contact Information	Required	Manual, then Automatic	R
dc:rights/Rights Statement	Required	Semi-automatic or Automatic	NR
dc:description/ Description of the Data Set	Optional	Manual	NR
dc:subject/Keywords Describing the Data Set	Required	Manual and Automatic	NR
dc:coverage / Locality	Required	Semi-automatic	R
dc:coverage/Date Range	Required	Semi-automatic	R
EML:software/Software	Optional	Semi-automatic	R
dc:format/File Format	Required	Automatic	NR
dc:format/File Size	Required	Automatic	NR
dc:date/(Hidden)	Required	Automatic	NR
dc:date/Date Modified	Required	Automatic	NR
Darwin Core: species/ Species, or Scientific Name	Optional	Semi-automatic	R

[1] "The DRIADE Project: Phased Application Profile Development in Support of Open Science", DC2007. (in press) http://www.dc2007.sg/

Summary / Future Work

Accomplishments

- Established prioritized goals, functional requirements, phased development plan, metadata application profile.
- Identified research priorities.

Next steps

- Refine DRIADE's functional requirements.
- Flesh out the phased development plan for the repository.
- Research plans:
  - Conduct two-part study to better understand evolutionary biologists' data use, preservation practices.
  - Evaluate controlled vocabularies.

Acknowledgments / Contact Information

We would like to also acknowledge contributions from DRIADE team members: Todd Vision, Associate Director of Informatics, NESCent, and Assistant Professor, UNC; and Hilmar Lapp, Assistant Director of Informatics, NESCent.
This work is supported by National Science Foundation Grant # EF-0423641.
Contact: jdube@email.unc.edu or visit: https://www.nescent.org/wg_digitaldata/

UNC School of Information and Library Science — Metadata Research Center (MRC) — http://ils.unc.edu/mrc/

National Science Foundation

NESCent National Evolutionary Synthesis Center — http://www.nescent.org/