PROJECT DESCRIPTION

The field of evolutionary biology is suffering from a crisis of data attrition. The problem is particularly evident when a researcher unsuccessfully attempts to obtain data sets associated with a published journal article. Though specialized databases exist for some of the most commonly seen data types (such as DNA sequences, character state matrices, and phylogenetic trees), 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 make their data accessible on-line, many individual researchers lack the technological means and sustainable infrastructure to ensure preservation and availability of their data over the long term. In this respect, evolutionary biology is typical of "small science" disciplines -- individuals or small groups collect much of the data manually, datasets are highly idiosyncratic in composition and format, and there is little infrastructure available for authors to share published data. As a result, much of the data underlying published works in the field is unavailable for future researchers to validate controversial findings, to reuse for studies that build upon the published work, to reanalyze as new methods and ideas are introduced, and to synthesize for the discovery of emergent trends. 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 throughout the discipline.

The **overall aim** in this proposal is to facilitate data sharing upon publication by the evolutionary community by addressing the major hurdles to adoption of Dryad, both technical and otherwise, in three broad areas: i) deposition and access interface, ii) incentives and interoperability, and iii) sustainability. We will also promote the use Dryad as an educational tool to teach future scientists about the value of digital data archives. To achieve these goals, we propose the following specific aims (SA).

1. Deposition and access interface. Dryad aspires to provide a way for researchers to deposit their data in a usable form with minimal burden, and to take fuller advantage of existing technologies for information retrieval. **[SA1.1]** Data deposition will be coordinated with the manuscript submission process. This will enable reliable bibliographic metadata (e.g. author, title, etc.) to be automatically stored by Dryad, and the citation for the data objects can be automatically included in the article. **[SA1.2]** We will explore means for assisting the capture of scientific metadata (e.g. geo-spatial information, taxonomic scope) from authors using various approaches in automatic metadata generation. **[SA1.3]** To maintain both data integrity and metadata quality, data curators will validate and, if necessary, edit, submissions to the repository. **[SA1.4]** A retrieval interface will be developed that uses the available metadata more fully, and also uses both existing and newly developed and relevant vocabularies to augment queries **[SA1.5]** Evaluations and user-testing will be employed during the design and implementation process, including studies of automated and user-generated metadata quality, the accuracy and recall of information retrieval, and usability studies of both the deposition and retrieval interface.

2. Incentives and interoperability. A major incentive to adoption is to implement, as far as possible, "one-stop shopping" for the deposition, discovery and retrieval of data. Towards this goal, we will enable interoperability with specialized databases and with metadata registries in related disciplines. **[SA2.1]** As proof-of-concept for one-stop deposition, we will implement hand-shaking mechanisms with GenBank, for sequence data, and TreeBASE, for phylogenetic data so that, where required by the journal or requested by the author, data will simultaneously be deposited in Dryad and either GenBank or TreeBASE. Handshaking will include automatic reuse of bibliographic metadata and identifiers, greatly simplifying the task of data deposition for the author. **[SA2.2]** Dryad will assign globally unique, stable, and resolvable identifiers for datasets. These identifiers will enable Dryad to broker among the data objects related to a single paper, whether they be within Dryad itself or in a specialized repository and data identifiers will provide a mechanism for data citations. **[SA2.3]** Interoperability of Dryad with other digital collections in biology and beyond will be achieved, in part, by implementing the OAI-PMH protocol for metadata harvesting. As a proof-of-concept, we will add full compliance with OAI-PMH to Dryad, TreeBASE and Metacat, the premier metadata registry and data repository for ecology. **[SA2.4]** Dryad and MetaCat will also implement the Library of Congress Search and Retrieve via URL standard, which will allow on-the-fly access to repository contents by third parties through a web-service protocol, and will also enable syndication of repository contents.

3. Sustainability. We propose a governance model and one technical experiment designed to ensure data preservation and sharing in perpetuity. **[SA3.1]** Dryad will be overseen by a Management 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 a distributed data preservation system following the LOCKSS (Lots of Copies Keep Stuff Safe) model, in addition to managing a more standard architecture of redundant production and backup systems within the North Carolina State University Libraries.

4. Community engagement is an integral component of the project and is critical both to short-term adoption by the user community and its long-term success. **[SA4.1]** Datasets of special educational value will receive extra curatorial attention and be presented for student use through a dedicated education section of the repository, acclimating future investigators to a scientific culture in which digitally shared data will play an increasingly important role. **[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]** NESCent will host annual workshops 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) and a pioneering application of digital library technology to data sharing for “small science”. We intend that this will serve as a model for efforts to preserve and share data in other disciplines facing a similar crisis of data attrition.

Results from Prior NSF Support.

K. Smith. NSF EF-0423641, "CSBE: A Place for Evolutionary Synthesis in North Carolina’s Research Triangle", $15,000,000, 12/1/04-11/30/09. This is the core funding for the National Evolutionary Synthesis Center (NESCent), which has funded administrative, outreach and informatics staff, over 20 resident postdoctoral and sabbatical scholars, and approximately 700 scientists participating in a variety of working groups and other meetings sponsored by the Center. NESCent's informatics group supports the sponsored science program of the center and spearheads cyberinfrastructure initiatives in open source database, software, and semantic web technologies for evolutionary biology. The education and outreach program translates the results of evolutionary biology to the education community and general public and helps recruit evolutionary biologists from underrepresented groups. **Other personnel.** (Note that while Duke University does not grant official co-PI status for subcontract recipients, the following personnel are co-PIs on this proposal in all other respects). **J. Greenberg**. NSF EF-0423641, UNC/Metadata Research Center, School of Information and Library Science, subcontract to NESCent core grant (see above), $133,417.87, 12/1/06-12/31/07. This work included background research and planning for the development of Dryad. Research accomplishments include a needs assessment examining the functionalities and features of selected data repositories and digital libraries; an analysis of selected controlled vocabularies and ontologies to determine their applicability for indexing data objects in and journal articles in Dryad, an empirical analysis of data underlying publications in evolutionary biology, and the development of a metadata application profile and cross walk analysis. Two publications to date ([1, 2]). **W. Michener**. NSF DBI-0225665, "ITR: Collaboration Research: Enabling the Science Environment for Ecological Knowledge (SEEK)", $4,439,765, 10/1/02-9/30/07. SEEK is a five year multi-institutional, multi-national initiative designed to create cyberinfrastructure for ecological, environmental, and biodiversity research and to educate the ecological community (especially, under-represented groups) about ecoinformatics. SEEK participants have designed an integrated data grid (EarthGrid) for accessing a wide variety of ecological and biodiversity data and analytical tools (including Kepler, an open-source scientific workflow solution). R. Waide, J. Brunt, W. Michener, J. Vande Castle. NSF DEB-0236154, "Network Office of the US LTER", $9,011,235, 3/1/03-2/28/09. The LTER Network Office provides administrative, cyberinfrastructure, training, and scientific synthesis support for the network of 26 Long Term Ecological Research Network sites that are located in the United States, Puerto Rice, French Polynesia, and Antarctica. **W. Piel.** EF 0331654: ITR: Building the Tree of Life — A National Resource for Phyloinformatics and Computational Phylogenetics (PI: T. Warnow) $293,245 subcontracted to Piel, 2003-2008. This collaborative project aims to establish a national computational resource to allow scientists to reconstruct the tree of life. Piel is part of the TreeBASE II team that has developed the next generation phylogenetic database at the San Diego Supercomputing Center. For publications see [3-6].

Introduction

Publication is at the heart of the scientific enterprise since it is principally through the peer-reviewed publication process that scientists make their findings known to their colleagues, the end-products of that process are viewed as uniquely authoritative, and such publications are the universal currency of professional achievement in science. By openly describing findings in a way that can be validated, repeated and built upon, science advances and scientists can lay claim to the credit for their work. Due to the premium placed on the body of respected and cited publications, the data underlying publications have particular value. The scientific field as a whole has a vested interest in the openness, availability and reusability of that data, an idea sometimes referred to as Open Data. The concept of Open Data is distinct and separable from that of Open Access, which is the idea that dissemination of the paper itself should not be restricted by its copyright. However, the two are similarly motivated by the principle that scientific products are public goods that are diminished by restricted access.

In an influential National Research Council report, Cech et al. [7] articulated what the committee called the **Uniform Principle for Sharing Integral Data and Materials Expeditiously** (UPSIDE): "*Community standards for sharing publication-related data and materials should flow from the general principle that the publication of scientific information is intended to move science forward. More specifically, the act of publishing is a quid pro quo in which authors receive credit and acknowledgment in exchange for disclosure of their scientific findings. An author's obligation is not only to release data and materials to enable others to verify or replicate published findings [...] but also to provide them in a form on which other scientists can build with further research. All members of the scientific community — whether working in academia, government, or a commercial enterprise — have equal responsibility for upholding community standards as participants in the publication system, and all should be equally able to derive benefits from it.*"

The NRC committee identified a number of additional corollary principles, including the following: "**Principle 1.** Authors should include in their publications the data, algorithms, or other information that is central or integral to the publication—that is, whatever is necessary to support the major claims of the paper and would enable one skilled in the art to verify or replicate the claims. **Principle 2.** If central or integral information cannot be included in the publication for practical reasons (for example, because a dataset is too large), it should be made freely [...] and readily accessible through other means (for example, on-line). Moreover, when necessary to enable further research, integral information should be made available in a form that enables it to be manipulated, analyzed, and combined with other scientific data. **Principle 3.** If publicly accessible repositories for data have been agreed on by a community of researchers and are in general use, the relevant data should be deposited in one of these repositories by the time of publication."

Open Data and Orphan Data

During the early to mid-20th century it was customary for extensive numerical tables, images, and other raw data to be published as part of a scientific article. Although the high costs associated with printed media have curtailed this practice in recent decades, digital technology has seen a resurgence of data sharing through the availability of supplementary material hosted on the websites of publishers and self-archived on the websites of individual labs. In addition, many journals now require certain types of data to be deposited prior to publication in an appropriate repository, such as GenBank (or EMBL, in the case of DNA sequences), PDB (for protein structural data), and TreeBASE (for phylogenetic data). By specializing on particular datatypes, these repositories can have highly structured data, rich metadata, and analytical capabilities uniquely tailored to their contents (e.g. 3D visualizations in the case of PDB). In evolutionary biology, only GenBank (and its international partners) and TreeBASE are typically mandated by journals. GenBank (and its European counterpart EMBL) is the preeminent sequence database. Treebase is a relational database of published phylogenetic trees and the data matrices used to generate them.

For the many orphan datasets for which specialized repositories do not (yet) exist (e.g. tables of numbers in a format unique to a given study) journals (and funding agencies) usually expect that authors will share it upon request. However, authors frequently decline or are unable to share their data upon request and few journals have incentives to encourage compliance. In a survey of 1240 geneticists, 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 [8]. Of the

12% that admitted to intentionally withholding data themselves, the most common reasons were: (1) that it required too much effort to produce the materials or information (80% of respondents); (2) that they were protecting the ability of a graduate student, postdoctoral fellow, or junior faculty member to publish (64%); or (3) that they were protecting their own ability to publish (57%).

A sobering example from psychology was described by Wicherts et al [9], who emailed requests for data from the corresponding authors of all 141 empirical articles that had been published in the most recent two issues of four American Psychological Association (APA) journals. Their aim was to reanalyze the data and assess the robustness of the research findings to outliers. All of the authors were presumed to have signed the APA Certification of Compliance with APA Ethical Principles, which includes the principle of sharing data for reanalysis. However, Wicherts et al. [9] report that "6 months later, after writing more than 400 e-mails–and sending some corresponding authors detailed descriptions of our study aims, approvals of our ethical committee, signed assurances not to share data with others, and even our full resumes-we ended up with a meager 38 positive reactions and the actual data sets from 64 studies (25.7% of the total number of 249 data sets). This means that 73% of the authors did not share their data." These examples illustrate the obvious advantages of publication-triggered deposition of data.

To determine the extent and kinds of data that may be at risk of loss in the field of evolutionary biology, we performed a survey of 27 randomly selected articles published within the last nine months in five major journals (American Naturalist, Evolution, Molecular Ecology, Molecular Biology and Evolution, and Systematic Biology). When the data or results in the articles were not described in the text itself, they were typically provided as figures (e.g., maps, graphs, diagrams, photographs, phylogenetic trees - an average of five per article), equations, or results tables (an average of two per article). Supplemental data was provided for 41% of the articles, and was typically composed of similar types of data objects as in the paper itself. Raw data tables were rarely provided by any of the articles, even in the supplemental data. Exceptions include a set of simulation results posted as supplemental data, and a number of papers in which sequence alignments available as supplemental data. However the majority of studies (67%) that made use of alignments did not make them available as supplemental materials. The vast majority of articles (78%) were based at least in part on the analysis of datasets not deposited in any repository. This is likely due to the unavailability of a suitable place for these highly variable types of data, which included biological measurements (e.g., morphology, life history, behavioral observations), genetic data sets, sequence alignments, gene annotations, and simulated data. Only 7% of the authors made these orphan datasets available by posting as supplemental data on the journal's website or by self-archiving. When a suitable data repository was available (e.g. GenBank, TreeBASE) data were generally deposited there. Interestingly, 48% of the articles based some or all of their conclusions on data from previously published studies. Thus, by at least this crude measure, authors are using shared data more frequently than they are sharing it themselves! Another important point is that if a dataset is not explicitly prepared for sharing in a digital form at the time of publication, it is at risk of having a short shelf life. Even with the best of intentions, data files are lost, become corrupted, or the proprietary software in which they were produced becomes obsolete, idiosyncratic data formats and coding schemes become increasingly incomprehensible as memories fade, and people move on. It is likely that many, if not most, of the orphan datasets underlying publications in evolutionary biology are irretrievably lost within a decade or two. This is a wasteful use of scarce human resources and research dollars.

Some journals have attempted to address this crisis. The Society of Systematic Biologists maintains an archive of the datasets underlying each publication in Systematic Biology on their website (www.systbio.org) which keeps the data package from each paper intact, offers much improved long-term preservation prospects over self-archiving, and provides a home for data types lacking an appropriate specialized repository. In principle, individual archives of this nature, one for each journal, publisher, or scientific society, would help address the crisis of data loss alluded to above, and would provide the means to replicate and validate the findings of individual publications. However, this model has its limits. Published data are far more useful for subsequent scientific discoveries when they are exposed to other data in a centralized repository, where software tools can be employed to search, visualize, analyze the contents of a digital data archive in ways the authors may not have imagined, and data from different studies can be combined in novel ways. Since the number of data types far exceeds the number of specialized repositories that are likely to emerge within the foreseeable future, a preservation repository would need to be a catch-all that can steward any type of digital data, be it tables of numbers, maps, photographs, sound files, etc. An additional advantage of a centralized repository is the economy of

scale that it allows. No single society or journal is sufficiently wealthy to adequately devote resources to a repository of this breadth for the indefinite future. Even for data that is "born digital", files must be migrated to new formats and media over time. Successful preservation of a digital archive requires the long-term stewardship perspective more commonly seen in the academic library setting. Thus, it makes both functional and economic sense to pool the resources of journals, societies, and other institutions discipline-wide, particularly in a relatively resource-poor discipline such as evolutionary biology.

The landscape of digital data repositories.

Recent years have seen the development of sophisticated technologies for digital data repositories [10]. Many of these have naturally arisen out of "large science" disciplines in which much of the primary data is systematically collected by a few data centers (e.g. observatories in the case of astronomy). A number of these initiatives have been in fields that closely intersect with evolutionary biology, such as the Knowledge Network for Biocomplexity (KNB), Science Environment for Ecological Knowledge (SEEK, see [11]), the Chronos consortium for geosciences (see [12]), and a variety of different efforts in the area of genomics. The Knowledge Network for Biocomplexity (see [13]) developed a suite of tools, including Metacat (an XML driven data repository, described more fully below), the Ecological Metadata Language (EML), which can be used to describe ecological data in a standard format, and Morpho, a desktop application for accessing and manipulating data and metadata both locally and over the network (see [14]). In other fields, the Dataverse Network is integrating web software, networking protocols, data standards, and analytical tools in a network of repositories for (primarily quantitative) social science research data [15].

However, relatively few efforts have thus far focused on enabling authors in small science disciplines to share data packages supporting publications. In the social sciences, there has been a movement towards deposition of “replication data sets” along with papers [16], and Systematic Biology has, for a number of years, maintained an archive of datafiles from all its published papers (see [17]). Other efforts attempt to link data with papers *post*-publication. One example is ChemXSeer ([18, 19]), an experimental system for processing the published chemistry literature and extracting data (e.g. tables) and metadata (e.g. chemical names) from papers without human intervention. The BioLit project (NSF BDI-0544575) is a recently funded pilot project to integrate the Public Library of Science (PLoS) journals with relevant entries in the Protein Data Bank (PDB).

In addition to these efforts within specific scientific disciplines, the information and library science community has long recognized the need for digital asset management by academic and instutional libraries [10]. Among the various tools developed with this mission in mind, one of the most successful is DSpace [20, 21], a freely-available, open-source archiving system designed by MIT Libraries and Hewlett Packard for the capture, management and sharing of digital assets. DSpace is by far the most popular software for institutional repositories [22] - over 200 institutions have adopted it - and it has a well-supported and active developer community. In 2006 the non-profit DSpace Federation was founded as a software development governing body (see [23]) with a long-term vision, and has since provided architecture and development goal roadmaps. Several recent developments are particularly relevant to this proposal. The Manakin project (see [24]) has created a modular and highly customizable user interface for searching, browsing, and display of digital assets, and the Configurable Submission System (see [25]) is working to make the deposition workflow highly customizable. Additional efforts are being invested in making DSpace agnostic of the actual identifier system and metadata schema. Thus, technologies are being developed within the information and library science community that are well-suited to the needs of a small-science digital data repository.

Community consensus on data sharing needs

NESCent and the Metadata Research Center hosted a workshop in May 2007 entitled “Challenges for Small Science Communities in the Digital Era" (see [26]). The forum brought together a variety of evolutionary biology stakeholders (i.e. journal editors and representatives from scientific societies) with experts from the realm of data centers, digital libraries, and data sharing. Workshop participants formulated specific recommendations in the areas of (1) adoption and sustainability, (2) intellectual property, (3) technology, and (4) digital asset lifecycle management. This workshop reinforced earlier feedback we had received from stakeholders in the evolutionary biology community that **immediate data preservation, staunching the flow of data loss, should be a very high priority** for NESCent. When pressed, participants consistently express a willingness to postpone higher Open Data goals (see Figure

1) in order to preserve, in some form, the large amount of publication-related data that would otherwise be lost.

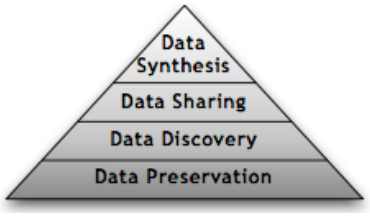

Figure 1. An informal hierarchy of Open Data goals. Preservation is the prerequisite to achieve the higher goals of resource discovery, sharing and synthesis. Synthesis, or data integration, requires substantially more detailed metadata than resource discovery and data sharing, and is therefore a more elusive and ambitious goal for a general data repository.

A major outcome of the Small Science Workshop was a Joint Data Archiving Policy, initially drafted by Michael Whitlock, Editor-in-Chief of The American Naturalist. The policy is intended to be adopted by a consortium of journals once a shared repository for data preservation is in place:

Joint Data Archiving Policy: *<<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, TreBASE, or Dryad [see below]. 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.*

Whitlock made the following additional points about the policy (quoted from the memo):

1. *The aim is for the consortium of journals to adopt the policy simultaneously, expressly to prevent any journal from being penalized by having an unusual policy. The cooperative nature of the policy is essential for its success.*
2. *Archived data will be citable along with the journal article itself so that authors receive fair attribution and credit for their work.*
3. *The policy requires only " the data ...with sufficient details that, together with the contents of the paper, allows each result in the published paper to be re-created." By the inclusion of this phrase, the policy intentionally does not require the author's entire data set to be archived, but rather the raw data involved only in that paper. Of course, it is hoped that authors will choose to archive larger, intact, data sets, but this policy is intended to preserve the authors' rights for subsequent uses of the data, while making that data that has been publicly described publicly available.*
4. *Journals may choose to allow an option for a one year embargo after publication, during which time the data are archived but not publicly available. Again, the option for no embargo is available for any author, but it is expected that this embargo option will smooth the road to acceptance by the broader community.*
5. *Some data should not be publicly available, such as the locations of endangered species populations. The consortium will not require, and indeed encourages blocking, archiving of such information. Other data from these studies should not be sensitive and can be archived.*

Although there is not yet consensus on all details, the idea of a joint deposition policy has been enthusiastically endorsed by the evolutionary biology community, as attested to by the letters of support from some of the most prestigious journals and scientific societies in the field (appended to this proposal).

Dryad: a digital repository for publication-related data

To help make the Joint Data Archiving Policy possible, NESCent has begun to develop a shared digital repository for data underlying published works in evolutionary biology. The system being developed is named Dryad, after the preternaturally long-lived tree spirits of Greek mythology. The development of Dryad is being staged in three phases:

Phase I consists of an initial repository implementation using a relatively standard implementation of DSpace. Phase I is being supported by NESCent core funding and will be underway before the start of this granting period. This will provide a means to preserve intact data packages from publications as soon as possible, and also serve as a useful testing environment, but it will lack many of the Phase II features that are deemed to be critical for wide acceptance by the research community.

Phase II is what is described in this proposal, and implements the primary recommendations that emerged from the Small Science Workshop. Dryad Phase II will integrate data deposition with publication and with specialized repositories, thus providing one-stop-deposition for authors. It will allow identifiers and metadata to be shared among the various digital representations of the publication and data package. Policy and strategic management will be guided by the stakeholders themselves. Data will be automatically and manually curated to ensure validity of the digital assets and thus their reusability. The deposition interface will be made as user-friendly as possible using automated metadata generation. Journals are not expected to require deposition of their authors until Dryad has achieved the goal of making deposition sufficiently smooth in the opinion of the Managing Board, though voluntary submission will be strongly encouraged before that time. It is intended that the interface development work will allow the Joint Data Archiving Policy to be adopted in Yr 3 of the current proposal.

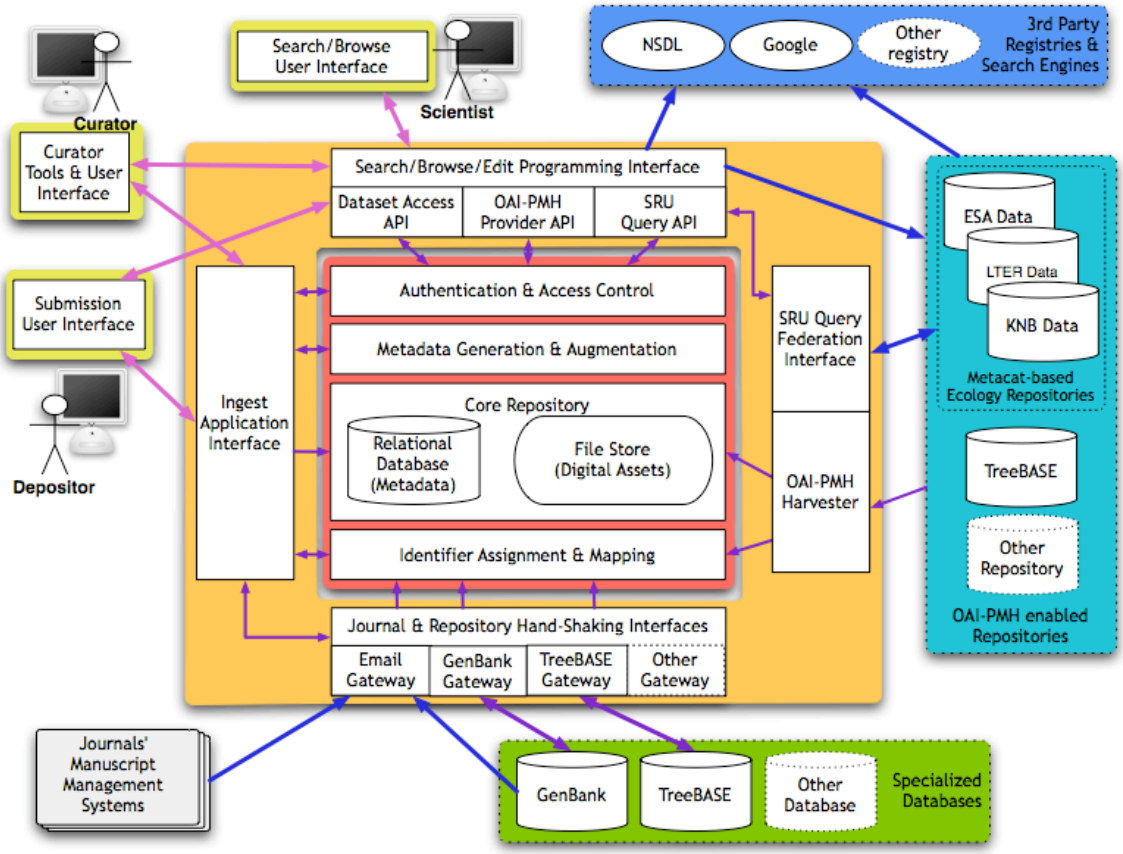

Figure 2. Overview of the proposed Phase II Dryad repository and its external interactions. See the Project Plan for a description of the compartments and their relationships to each other.

Phase III is planned for after the completion of the proposed work and once Dryad has become financially self-sustaining. It will take fuller advantage of the centralized collection of data packages with a host of implicit and explicit relationships to each other by featuring more extensive data integration, means for capturing user knowledge, and integrated analysis capabilities. Since the immediate concerns of the journal consortium are data preservation and community adoption, the current proposal is focused primarily on the Phase II aims.

Project plan

1. Deposition, retrieval and data curation. As currently envisioned, data are to be deposited in Dryad once an article has been accepted, rather than prior to peer review (although this is a policy decision which may change over time or differ among journals). Phase II aims to make deposition easy for authors through coordinated submission of the manuscript and associated data package, and to make a rich data

discovery interface by innovative use of metadata and vocabularies.

1.1. Coordination with journal submission. Almost all journals employ a manuscript management software system that assigns an identifier and captures bibliographic metadata (e.g. authors, title, abstract, subject keywords) at the time of submission. These same bibliographic metadata are among the most important for Dryad to capture in order to unambiguously attribute each dataset to the correct article, authors, etc. Manuscript management software systems regularly employ customizable email form letters to communicate status changes and information requests to the various parties involved in manuscript processing. Dryad will take advantage of this fact to implement a robust three-way communication system among the journal, the author/depositor, and the repository that involves no more work or dependencies on the side of the journal than sending an electronic acceptance letter.

Bibliographic metadata recorded by the manuscript management system for each submitted manuscript will be transmitted to Dryad through email form letters and parsed using an electronic gateway that extracts and stores the bibliographic metadata. The system will also generate an electronic email notice to the corresponding author, providing him or her with a URL and a randomly generated authorization token to register with the repository and deposit the associated datasets for the manuscript. In this way, Dryad already is in possession of all bibliographic metadata at the time an author accesses the Dryad submission interface; the metadata will merely need to be verified by the depositor. Typically these metadata will be highly accurate since both the manuscript author as well as the journal editor will have already reviewed them. Upon completion of data deposition in Dryad, the assigned data identifiers (SA2.3) will be relayed back to the journals to enable augmentation of the published article with data citations (SA2.3). Deposition and modification of data will require both authentication and authorization. To limit proliferation of accounts and to encourage more secure credentials we will investigate possibilities for employing user-centric digital identity schemes such as OpenID (see [27]). Once deposited and assigned an identifier (see below), modifications to data objects trigger a new version, which will also receive a new identifier, while. older versions will remain accessible.

The specification of what data authors will be required to deposit is left as a policy decision for journals, including questions of whether to include software, simulation results, unprocessed data, etc. Similarly, enforcement of the data deposition policy will be the responsibility of the journal.

1.2 Metadata capture. Accurate and extensive metadata is needed for digital resources to be discovered and properly used, but it is not easily obtained. Projects that rely on data centers for primary data gathering typically use metadata schemes that are far more elaborate than would be appropriate for data depositors from the general scientific community. For example, the Ecological Metadata Language includes over 100 properties and subproperties combined, making the generation of an accurate and comprehensive EML description a daunting prospect for an individual researcher. On the other end of the scale, it is also problematic when metadata is so sparse that it limits resource discovery and reuse. For example, one can search the supplementary data archives of *Molecular Biology and Evolution* via author, title, publication year, and volume number, but not by gene/protein, taxa, data format, or other aspects of the data that could be extremely useful for retrieval.

Dryad will use a relatively lightweight metadata application profile that draws elements from a variety of schemes, including the Dublin Core, the Data Documentation Initiative (DDI), Darwin Core, EML and PREMIS (see [2] and references therein). The profile consists of a *bibliographic citation module*, which holds information that is obtained through the manuscript submission interface, and a *data object module*, which holds metadata that is, for the most part, automatically generated by the system (including 16 of 18 elements in the preliminary profile, such as file format and deposition date), as well as some metadata specific to each data object that needs to be supplied, at least in part, by the depositor. There may be one or more such data objects associated with each bibliographic citation. To minimize burden on the depositor, required metadata for each object will be very limited (e.g. a brief free-text description). At the same time, the interface will contain optional fields that can be used to capture richer data (such as taxonomic and geographic scope) from depositors who are motivated to do so. We will also explore incentives, such as having a featured datasets on the journal and/or repository home pages, in order to motivate authors to provide additional metadata (using elements drawn largely from the EML scheme: taxonomic coverage, geographic coverage, etc.). Furthermore, Dryad will capture more extensive metadata during the process of handshaking with specialized databases (SA2.1).

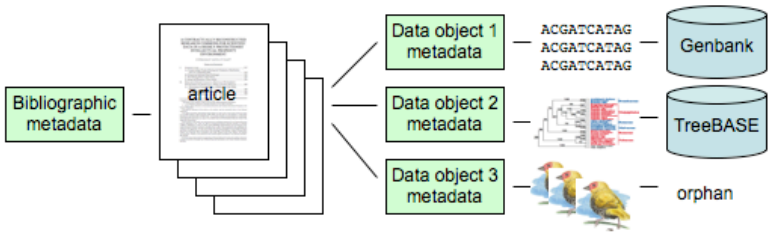

Figure 3. A hypothetical published data package. In this example, sequence data is shared with GenBank, phylogenetic data is shared with TreeBASE, and orphan data (represented by bird images) is captured only by Dryad. Metadata is associated both with the article and with each data object. The article itself is used by Dryad, but not exposed.

Another promising way to obtain more extensive metadata without relying entirely on depositor effort is to employ automated systems for metadata generation and extraction [28-30]. Automatic approaches have been shown to be more efficient, less costly, and more consistent compared to human approaches [31]. In experimental settings, automatic metadata generation that employ machine-learning and natural language processing techniques have been successful with semi-structured data as input (e.g., [32, 33]). The standard structure and format of journal publications suggests that automated metadata generation will also be successful in the context of Dryad.

Such methods are particularly promising in conjunction with the use of ontologies, thesauri, and classification systems (e.g., [34, 35]). We conducted a preliminary analysis to determine the coverage of existing vocabularies for keywords used by scientists in the field by sampling ~600 author-identified keywords (or phrases) from 140 articles (from the same sample of journals as described above). Only 16% were duplicates, indicating surprisingly limited agreement among scientists in their use of vocabularies and suggests that controlled vocabularies could have a major impact on indexing and information retrieval. The 551 unique terms were manually classified into one of eight facets: concept, field/discipline, gene, habitat, method, place name, taxon name, and time period. Terms from each facet were searched in a standard set of relevant vocabularies. The highest match rate (72%) was achieved for taxon names, using ITIS [36] and the uBio NameBank [37]. No other facet had a match rate greater than 50%, another indication of the need for development of a vocabulary specifically for Dryad.

To examine automated metadata generation and build a controlled vocabulary suitable for this task, we will use an extensive corpus of articles from the consortium journals. We will test known automatic indexing algorithms and methods, including the Lemur toollkit (see [38]) and Clairlib (see [39]). Automatic extraction and controlled vocabulary term matching processes will be used for assigning metadata values, which can then be verified or augmented by the user or the data curator (SA1.3). We will study these techniques particularly for the semi-automatic generation of a title and description for each data object (since this must be supplied and cannot be extracted from the manuscript submission system). In addition, we hypothesize that automatically generated metadata will greatly assist the author by shortening deposition time and encouraging them to provide more extensive and higher quality metadata; we propose to test this hypothesis under SA1.5.

1.3 Data curation. To maintain both data integrity and metadata quality, Dryad will be staffed by data curators, whose responsibilities will include validation of the formats of the deposited files, editing metadata where necessary, communication with authors/journals when problems arise, helping to verify the usability of metadata, overseeing data format migration, and serving as a help desk for depositors, and presenting tutorials on the use of the repository at the annual meetings of the consortium societies (SA4.2). Curation will be assisted by custom software for metadata quality assessment (see below), as well as existing software such as JHOVE and Xena (see [40-42]) for format validation and migration. We will study and incorporate methods for automatic measurement of metadata quality by drawing on and extending work of the AMeGA [28, 29] and Infomine projects [43]. Dryad will provide an empirical measure of metadata quality for each metadata record using a variety of metrics (for instance, the match rate between a document and a controlled vocabulary). The rating will help the curator determine which metadata records require review and whether the original depositor needs to be contacted. The curators will also target a limited number datasets of special data packages (i.e. those that are frequently downloaded by users, or those that are particularly suitable for educational purposes) for a higher-level of

curatorial attention (SA4.1). Finally, the curators will participate in the metadata generation and quality evaluation studies (SA1.2 and SA1.5). Curators will not be expected to validate the biological correctness of the data itself, or to determine the completeness of each data package. In this proposal, we have allocated resources for two half-time, postdoctoral-level, curators.

1.4 Data retrieval. Standard journal data archives (e.g. Systematic Biology) allow users to retrieve data packages once they have navigated to a specific bibliographic record. Building on the metadata resources described under SA1.2, much richer retrieval functionality will be provided by Dryad. For instance, aside from typical author, title and keyword searches, users will be able to search for data records based on facets such as taxonomy, geography, research method, time period, etc. Queries will be expanded based on available ontologies and thesauri, including the vocabulary being developed for this project, to include hierarchical, associative, and equivalent relationships among potential search terms (SA1.2). For example, taxonomic searches will automatically search descendents within the taxonomic hierarchy. We will conduct studies of Dryad's retrieval precision and recall, as well as study the information-seeking behavior of researchers to assist in the design of the retrieval interface.

1.5 Evaluation. Evaluations of Dryad will be conducted throughout the funding period and will be used to guide modifications to the system. We will conduct studies to compare the depositor-time requirements and quality of user-generated and semi-automated metadata. We will also measure the effects of metadata generation, classification, quality assessment and search methods on the accuracy and recall of information retrieval. Vocabularies will be tested by measuring the frequency with which user-generated and text-extracted terms are matched. Usability studies (including use-logging, profiling and follow-up surveys) will be conducted of both the deposition and retrieval interface, using the Interactive Design Lab at the UNC School of Information Sciences. Feedback from the evolutionary biology community will be solicited through general surveys, Q&A sessions at society annual meetings, through logging of email and website comments, and by follow-up interviewers with depositors. If deemed necessary, the Management Board (see below) may contract an independent evaluation during Yr 3 for the purpose of determining whether the system is sufficiently operational for the Joint Data Archiving Policy to be put into effect.

2. Incentives and interoperability. The proposed work will promote adoption of Dryad within the evolutionary biology community through the visible and transformative impact it will have on the deposition process for authors, on the ability of users to search across diverse data collections, and on the level of connectivity among data objects that will be enabled by the use of global identifiers. It is worth noting that the infrastructure developed in the following specific aims can be used to achieve interoperability with a much wider community of existing and future specialized databases and digital resource-discovery systems than those that are specifically mentioned by name in this proposal.

2.1 Handshaking. Authors are already required by most journals to deposit newly generated nucleotide and protein sequence data into GenBank. Some journals, such as Systematic Biology, Molecular Phylogenetics and Evolution, Systematic Botany, and others, recommend or require submission of phylogenetic data to TreeBASE. There are other data types, ranging from fossil data to RNA expression profiling data, that may conceivably be part of the data supporting a published evolutionary biology article, and for which submission to a specialized database may be required by a journal. The requirement for data deposition in an increasing number of specialized databases can potentially place a large burden on authors who have to not only familiarize themselves with different submission interfaces, but also repeatedly enter the same metadata. This frustrates compliance with well-intentioned policies. Specialized databases augment the data through presentation and analysis capabilities specifically tailored to the respective type of data, and so there is no replacement for these resources. However, better models for getting the data into them deserve to be explored.

To do this, and as one of the key incentives for researchers to deposit data into Dryad, we propose to develop the submission interface of Dryad as a one-stop data submission tool capable of automatically submitting the individual data objects within a package to one or more different specialized databases. As proof-of-concept, we will design and implement a hand-shaking mechanism with TreeBASE and GenBank for automatically submitting phylogenetic trees/character state matrices, and nucleotide/protein sequences, respectively. These two repositories account, at present, for most (though not all) submissions by authors into specialized repositories mandated by consortium journals.

Dryad will collect any metadata required by the target database that has not already been captured, submit the pertinent data to the target database using a non-interactive programmatic gateway, and obtain the submission status, accession numbers, or possible error messages from the target database. This mechanism has the major advantage of facilitating the exchange of identifiers (SA2.3) between Dryad and the specialized databases, thus providing an electronic tether connecting the related data objects and their copies across different repositories and databases.

For TreeBASE, we will design and implement a robust, web-service based submission Application Programming Interface (API). An extensive redesign of TreeBASE by the CIPRES project (www.phylo.org) is scheduled for release in 2007. However, it currently lacks a submission API. The software to be added will include the automated data validation steps that are part of the new TreeBASE submission process (e.g. validating the NEXUS format, matching terminal taxa against the uBio NameBank). When TreeBASE rejects a submission, the depositor will be notified, advised how to correct the problem, and asked to resubmit.

To seed the repository with content early on, we will also populate Dryad with the published data presently in TreeBASE (currently ~1600 records). To provide meaningful metadata, we will extract from the legacy records such things as the original bibliographic citation and DOI, sequence accession numbers, geographic coordinates and specimen identifiers. For those articles that are also available electronically, this will result in an initial test set for automated metadata extraction techniques (SA1.2). We will also explore the possibility of similarly pre-populating Dryad with complete publication data packages from existing journal data archives.

For the submission of nucleotide and protein sequences or multiple sequence alignments to GenBank, a semi-automated process already exists. The tbl2asn from the National Center for Biotechnology Information (which hosts GenBank) merges feature and annotation tables with collections of sequences to generate files that can then be submitted to GenBank through an email or FTP gateway. GenBank staff review all submissions and, if accepted, send accession numbers to the submitting researcher. As part of a recently launched effort to further streamline this process to better meet the sequence submission requirements of the Barcode of Life project, NCBI staff will collaborate with us in automating submission through Dryad (see letter). We will create a tool using the NCBI C++ toolkit that non-interactively creates the file with the required metadata, almost all of which will be known to Dryad already. GenBank staff will send assigned accession numbers (or the error report) both to the depositor and to a Dryad gateway, and the metadata from GenBank will then become part of the Dryad record. GenBank staff will also identify a sequence annotation element through which the global identifier of the data submission (SA2.3) can become part of the GenBank record. This will enable GenBank users to easily retrieve all related data objects (such as the phylogenetic tree for which the sequence was used) through Dryad and any other site that allows for a search of the global identifier (such as TreeBASE).

In addition to these handshaking partners, Dryad will organize workshops (SA4.3) with representatives from other specialized databases in the field, in order to plan for handshaking mechanisms that would allow for authors to deposit data, through Dryad, that are not currently required by consortium journals (e.g. image data from comparative anatomy studies)

2.2 Globally unique, stable, resolvable identifiers for data objects. In order to foster widespread adoption of data sharing and reuse, a crucial challenge to be met is to establish a standard for attributing data to the authors as much as for other scholarly works. The Joint Data Archiving Policy stipulates that the users of published data should cite the original article, and stakeholders have indicated that providing a mechanism to conspicuously credit the creators of the data is a high priority. Proposed standards for data citations (e.g. [44]) also suggest the inclusion of one or more unique identifiers that can be used to access the data itself. If implemented properly, such an identifier will enable the bibliographic citation for a given dataset to be easily recovered, even if digital representations of part or all of a given data object become widely scattered among different repositories.

Dryad will assign identifiers to data objects that are globally unique, stable (to changes in location or access protocol), and resolvable (cf. [45]). By “resolvable”, we mean that there must be a mechanism to unambiguously return the electronic location of the dataset using the identifier. Ideally, there is also a straightforward way to convert an identifier into a unique and resolvable URL, and a URL-based mechanism for obtaining metadata for any given identifier (e.g. OpenURL, [46, 47]). Dryad will implement an appropriate identifier system based on its technical merits, costs, and prospects for long-term support.

There are three identifier systems that meet the minimal requirements to various degrees and widely used within the biological, library, and publishing communities: Life Science Identifiers (LSID, [48]), CNRI's Handle System (see [49]), and DOIs (Digital Object Identifiers, see [50]). LSID is a Uniform Resource Name (URN) based standard that is popular among major resource and data integration projects, such as caBIO (see [51]) and BioMOBY [52]. Resolution services are still not widely known, though, and the LSID resolution standard lacks a provision for how to obtain metadata for the referenced object. Any organization may become an LSID-assigning authority, and the cost per LSID is negligible. CNRI's Handle System and DOI are closely related. Identifiers consist of a string of printable characters with a prefix that denotes the "Naming Authority", followed by a slash ('/') and a local identifier assigned by the naming authority. Both the Handle System and the DOI system specify a central resolution service, and handles and DOIs can be trivially converted into resolvable URLs by prefixing them with http://hdl.handle.net and http://dx.doi.org, respectively. There is a nominal fee (currently $50 annually) to register as a naming authority for CNRI and no additional costs are incurred to assign handles. For this reason, handles are popular with local and institutional repositories. DSpace, for example, uses CNRI Handles to name and link to its document holdings. DOIs are supported by the International DOI Foundation (IDF), an organization founded by members of the publishing industry. Though technically based on the Handle System, the DOI standard adds an administrative framework that ensures and promotes common practices among participating member organizations. An organization wishing to assign DOIs needs to find an appropriate Registration Agency, which must be an IDF member, and chooses its own business model and intellectual property scope. For journals and most publishers, the Registration Agency is CrossRef, whose business model and contractual terms are geared towards (oftentimes commercial and for-profit) publishers. CrossRef also operates an OpenURL resolver for all DOIs that it assigns. DOIs have been widely adopted in the publishing community, and are assigned to most published articles. Consequently, the resolution mechanism for DOIs is widely known even among end users. However, until recently a Registration Agency with a business model geared towards academic data providers did not exist, making cost a major obstacle in assigning DOIs to data rather than articles. Beginning in 2003, a consortium of German scientific organizations, universities, and libraries launched an initiative to create the infrastructure through which datasets can be deposited in an open archive, and become citable through assignment of a DOI free of charge to the depositor. As a result of this initiative, the German National Library of Science and Technology (TIB) in 2005 became an IDF-accredited DOI Registration Agency (see [53]) for scientific data worldwide. The actual assignment of DOIs is delegated to Publication Agents, which must comply with a certain set of requirements, such as long-term availability of deposited data, link to a publication, and quality-controlled metadata. We will investigate with representatives from TIB the terms under which Dryad could be designated as a Publication Agent with a feasible cost model. Initial discussions would put the annual cost at under $4000, assuming around 1000 published articles per year, with 10 datasets per publication on average (see letter of support).

2.3 Metadata exchange through harvesting. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, see [54]) is a standard for automatically harvesting metadata from digital repositories. It has been widely adopted by numerous repositories and data providers, including the National Science Digital Library (NSDL, [55]), BioOne ([56]) and Public Library of Science (PLoS, [57]), and even Google, which can use a registered OAI-PMH gateway to crawl (harvest) a repository's publicly available content. DSpace includes an OAI-PMH gateway that is based on OAICat (see [58]), an adaptable open-source implementation.There is also an open-source OAI-PMH harvester implementation that is compatible with DSpace (see [59]).

Metacat ("metadata catalog", see [60]) is a framework for storing, querying, and retrieving XML documents, particularly EML documents [61, 62], and is used by many of the major data registries and repositories in ecology, including the Ecological Society of America data repository (ESA, [63]), the Long Term Ecological Research Network (LTER, [64]), and others (see listing at [65]). There is a specialized protocol for synchronizing content among Metacat archives, but communication with non-Metacat repositories is currently not well-supported. TreeBASE is similarly invisible to metadata indexing by third parties. Given that there is considerable overlap between the disciplines of ecology and evolution, and it would be very desirable for users to be able to search throughout all these data collections through whatever interface is most familiar to the user, be it Dryad, MetaCat or even Google..

To achieve this, we propose to implement fully OAI-PMH-compliant gateways in Dryad, TreeBASE

and Metacat. Use of OAI-PMH will expose the contents of all three not only to each other, but also to internet search engines, bibliographic metadata registries, etc. OAI-PMH requires Dublin Core as a minimum metadata exchange standard, but does allow for additional metadata formats that a client may choose if the server supports it. Since Metacat stores metadata natively in EML, and Dryad uses custom application profile, cross-walks to Dublin Core format will be required to support minimal metadata harvesting by third parties. In addition, metadata records in MetaCat will be exposed in native EML for harvesters that are EML-aware. Since TreeBASE stores its metadata in a relational data model and not as XML documents, a cross-walk from a native XML format to DC is not needed; instead, the DC elements will be drawn from the metadata attributes stored in the TreeBASE relational database. In the OAI-PMH specification, repositories may also provide a resolvable identifier to the data object itself, and Dryad and MetaCat will take advantage of this feature so that harvesters may choose to access the data object either as a part of the harvesting process or on-demand from a user. Dryad will regularly harvest metadata records from TreeBASE and Metacat, allowing users to extend searches to those records.

2.4 Enabling on-demand queries and continuous feeds through web services. In addition to the OAI-PMH gateway described above, we propose to adopt and implement the Library of Congress SRU/SRW standard (see [66]) to enable third-party queries of both Dryad and Metacat. SRU/SRW stands for Search and Retrieve via URL (or Web-service) and is a standard internet search and retrieve protocol that is independent of technology platform, database implementation, and data/metadata format. The results of an SRU query are returned by the server in XML format, which a client may then parse and present to the user in a customized manner. SRW is a flavor of SRU that uses SOAP (see [67]) for server-client communication. Various freely available tools and libraries exist for implementation of SRU and SRW, including support for the full-text indexing engine used within DSpace, and a query language specification called CQL.

Over the last several years, standards have been developed for websites to broadcast structured content changes, such as news items, through electronic "feeds." The most popular such standards are the RSS (for 'Really Simple Syndication', [68]) and Atom XML (see [69]) formats. External parties can monitor such feeds automatically, aggregate the content from multiple feeds, and customize their presentation. Thus, feeds can be used for "mashing up" the content of one site with that of another. The recent introduction of Yahoo Pipes (see [70]) takes feed aggregation to another level by allowing a user to visually create pipelines of feeds, filters, and processing components, effectively enabling ordinary users to program the network of web-feeds. The SRU standard is currently being extended to allow RSS as a metadata response format. Dryad will implement this extension, once it is approved by the SRU standards body, in order to broadcast all newly added and modified datasets. The metadata elements encompassed by RSS are intentionally generic; in essence a feed will consist of one or more channels, with associated metadata such as title, description, date, etc, and one or more items such as author, title, description, identifier, and source URL.

3. Sustainability. In order to ensure the long-term viability of a community-owned resource such as Dryad, it is important to consider issues of governance and financial management. In addition, technical aspects of reliability need to be addressed.

3.1 Management. A management board (MB) comprised of stakeholders will have ultimate responsibility for the management of Dryad. The MB will generally be responsible for setting policy for Dryad (e.g. whether and how long to allow for data embargos after publication in special instances), and setting long-term strategic goals (e.g. what additional specialized databases require handshaking). One representative appointed by each of the consortium journals, or its governing society, will be included on the MB. The senior personnel on this grant will also serve as *ex officio* members. The consortium journals have agreed to appoint a member to serve on the board, which will meet annually. The MB may, in addition, appoint a smaller executive committee of rotating members who can meet more frequently and consider proposals for adoption by the board. The MB is also expected to invite relevant experts to attend the annual meetings to advise on issues such as intellectual property, sustainability, digital library standards etc., as well as representatives from data sharing initiatives in other disciplines. The MB will have the authority to invite and accept requests from additional journals to join the consortium, provided they have the same rights and responsibilities as existing members.

As necessary, consultants will be contracted to assist the MB. One need is to develop a business

plan for the financial sustainability of the repository beyond the granting period. Initial discussions have already taken place with stakeholders regarding the relative merits of various cost-return models, and it is understood that some combination of funds from society budgets, journal page charges, and other revenue streams will need to be established by the end of the project period in order to manage the repository over the long-term. Another issue is the policy regarding intellectual property for data objects. The Creative Commons and Science Commons projects (see [71, 72]) have developed legal instruments that empower authors to grant rights to the uses of their digitally available data, but the legal landscape is somewhat complicated by differences in intellectual property laws among countries, and so resolution of licensing policy will be one of the initial issues before the MB. The MB will also be responsible for negotiating the details of the Joint Data Archiving Policy, and for determining when it is appropriate for the policy to take effect.

3.2 Data storage and redundancy. To incorporate a focus on long-term preservation in the operation of Dryad from the outset, the production environment of Dryad, including the repository software and the asset store, will be hosted by the Digital Library Initiatives (DLI) group at North Carolina State University (NCSU). DLI has substantial experience building and hosting digital data collections and is a partner in the Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP).

Libraries have addressed the problem of preserving the content from digitally published journals, even when a publisher stops providing that content, or when the publisher goes out of business, by adopting a software tool and preservation model called LOCKSS (Lots Of Copies Keep Stuff Safe; [73]), a software tool and preservation model. (A variant of this approach that utilizes a controlled network is called Controlled LOCKSS, or CLOCKSS [74].) LOCKSS is a low-cost-of-entry mechanism that allows replication of digital assets among geographically disparate institutions using inexpensive hardware components. Digital content preserved through LOCKSS is highly resistant to local disaster, organized attacks or other tampering attempts, and to "bit-rot." The operating system and all software is loaded from write-protected media and therefore will return to its trusted state upon reboot. In order to permanently damage the content holdings, copies kept at multiple participating institutions would need to be damaged. Recently, libraries have begun to apply this model to data, as opposed to journal content. The MetaArchive Project (see [75]), a collaboration involving the Library of Congress, has successfully implemented a private LOCKSS network, similar to CLOCKSS, to preserve at-risk digital Southern Cultural Heritage data collections at six participating universities.

We propose to establish an experimental implementation of LOCKSS for preserving the data and metadata objects held in Dryad. To do this, we will develop a LOCKSS plugin that will implement Dryad-specific mechanism, rules, and policies for harvesting (crawling) content. Through registration with the LOCKSS Alliance, the plugin will be disseminated to participating libraries in future updates of the software. We will designate three initial LOCKSS installations for preserving content stored in Dryad to be located at NCSU, UNC, and NESCent, and four additional installations at geographically remote institutions (to be determined). This experimental system will complement a more standard system of multiple levels of storage redundancy and backup that will be implemented within the NCSU Library.

4. Community engagement. Integral to the project are activities that involve communication with the broad array of current and future researchers who will be both depositors and consumers of the data, as well as with other informatics projects that contribute to the digital data landscape of evolutionary biology.

4.1 Educational outreach. With the goal of familiarizing future evolutionary biologists with the concept of data sharing and its associated technology, we will develop a section of Dryad as a resource for educators. DryEd will contain specially prepared datasets designed for student investigations into different aspects of evolutionary biology. Data curators will select a limited number of datasets (1-2 per year) to receive extra curatorial attention, based on popularity or thematic area. Preference will be given to datasets likely to have strong resonance with students (on topics such as the evolution of antibiotic resistance or viral pathogenicity, domestication of companion animals, human origins, origin of life, etc). Curators will work with authors, and with the NESCent Education and Outreach Group, to provide detailed metadata, more extensive background and related material, and a set of suggested exercises appropriate for each dataset. Resources will be targeted at the Advanced Placement, college, and graduate levels. Multiple routes of dissemination will be pursued through the resources of NESCent's EOG group (see [76]).

4.2 Tutorials. Dryad tutorials, designed for active investigators in the field, will be prepared by the data curators with the assistance of other project personnel, and presented at the scientific conferences deemed most appropriate by the MB (2-3 conferences/yr). The aim of the tutorials will be to explain the role of NESCent relative to the journals and specialized databases, to demonstrate the deposition and retrieval interface, and to assist authors in increasing the extent and quality of the metadata provided by raising their awareness of metadata in general.

4.3 Workshops. Annual workshops of 10 or more participants are planned at NESCent which will bring together experts on particular metadata and interoperability standards in order to plan for future handshaking activities with specialized databases and related initiatives. The MB will assist in selection of workshop themes and participants.

Management plan. A timeline for the proposed activities is given in Table 1, which also shows the breakdown of responsibility by institution. The overall coordination will be under the direction of the Director of NESCent (K. Smith) and the Associate Director for Informatics (T. Vision). There will be annual all-hands project meetings at rotating locations, in addition to short-term visits by project personnel among project locations to coordinate installation of handshaking mechanisms, OAI-PMH and SRU/SRW implementations, and to conduct usability tests.

Table 1. The approximate duration of effort (shaded) for each Specific Aim, and the parties responsible. Some aims (e.g. SA1.1) are deemed to be completed once operational, even though the outcomes will be used throughout the life of the project.

SA	activity	responsibility	yr1		yr2		yr3	
1.1	coord. with journal submission	NESCent						
1.2	metadata capture & research	UNC						
1.3	data curation (TreeBASE)	Yale						
1.3	data curation (Dryad)	NESCent/UNC						
1.4	data retrieval	NESCent						
1.5	evaluation	NESCent/UNC						
2.1	handshaking	NESCent/Yale						
2.2	identifiers	NESCent						
2.3	metadata harvesting	NESCent/Yale/UNM						
2.4	web services	NESCent/UNM						
3.1	management board	NESCent						
3.2	data storage	NESCent/NCSU						
4.1	Educational outreach	NESCent						
4.2	Tutorials	NESCent						
4.3	Workshops	NESCent						

Software Engineering and Dissemination. Software development at NESCent will take place under the direction of Hilmar Lapp, the Assistant Director for Informatics (see Bio. Sketch). Dr. Ryan Scherle will be responsible for architecture, design, and implementation of the data repository and its surrounding software components. Dr. Scherle holds a PhD in Computer Science from Indiana University and has extensive experience in digital library projects and initiatives. Additional junior programming staff and short-term contractors will report to Dr. Scherle. The software development methodology will follow agile development principles whenever possible, with milestones being driven by use-cases largely corresponding to the specific aims. User and programming interface milestones will be iteratively developed, starting with functional prototypes, in order to elicit rapid and regular feedback from collaborators, stakeholders, and beta-testers. All user interfaces will be subjected to usability testing to optimize ease-of-use (SA1.5). All source code will be made open-source and hosted on SourceForge.net or, as appropriate, other open-source hosting sites, and therefore freely available. Changes to the DSpace or other open-source repository codebases will be coordinated with the respective development team so that they may become registered software patches with eventual full integration into the respective codebases.

References

1. Dube, J., S. Carrier, and J. Greenberg, *DRIADE: a data repository for evolutionary biology*, in *Proceedings of the 2007 conference on Digital libraries*. 2007, ACM Press: Vancouver, BC, Canada. p. 481.
2. Carrier, S., J. Dube, and J. Greenberg, *The DRIADE Project: Phased Application Profile Development in Support of Open Science*, in *DC2007 - Application Profiles: Theory and Practice. International Conference on Dublin Core and Metadata Applications*. 2007: Singapore.
3. Piel, W., M. Sanderson, and M. Donoghue, *Data mining in phyloinformatics: the small-world dynamics of tree networks.* Bioinformatics, 2003. **19**(9): p. 1162-1168.
4. Nakhleh, L., et al., *Requirements of Phylogenetic Databases*, in *International Symposium on BioInformatics and BioEngineering*. 2003. p. 141-148.
5. Wang, J.T., et al., *TreeRank: A Similarity Measure for Nearest Neighbor Searching in Phylogenetic Databases*, in *Proceedings of the 15th International Conference on Scientific and Statistical Database Management (SSDBM 2003)*. 2003: Cambridge, MA. p. 171-180.
6. Herbert, K., et al., *Lineage path integration for phylogenetic resources*, in *Proceedings of the 17th International Conference on Scientific and Statistical Database Management*. 2005: Santa Barbara, California. p. 117-120.
7. Cech, T., et al., *Sharing publication-related data and materials: responsibilities of authorship in the life sciences.* Plant physiology, 2003. **132**(1): p. 19-24.
8. Campbell, E.G., et al., *Data withholding in academic genetics: evidence from a national survey.* Jama, 2002. **287**(4): p. 473-80.
9. Wicherts, J., et al., *The poor availability of psychological research data for reanalysis.* American Psychologist, 2006. **61**(7): p. 726–728.
10. ARL-NSF, *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering - A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe* 2006: Arlington, VA.
11. *Science Environment for Ecological Knowledge (SEEK)*. URL: http://seek.ecoinformatics.org/ [accessed: 7/6/2007]
12. *Chronos consortium for geosciences*. URL: http://www.chronos.org/ [accessed: 7/6/2007]
13. *The Knowledge Network for Biocomplexity (KNB)*. URL: http://knb.ecoinformatics.org/ [accessed: 7/6/2007]
14. *Morpho*. URL: http://knb.ecoinformatics.org/morphoportal.jsp [accessed: 7/6/2007]
15. King, G., *An Introduction to the Dataverse Network as an Infrastructure for Data Sharing.* Sociological Methods and Research, in press.
16. King, G., *Replication, Replication.* PS: Political Science and Politics, 1995. **28**(3): p. 443-452.
17. *The Society of Systemic Biologists, and home of Systematic Biology*. URL: http://www.systbio.org/ [accessed:
18. Bolelli, L., et al., *ChemXSeer: A Chemistry Web Portal for Scientific Literature and Datasets*, in *Open Repositories Conference*. 2007: San Antonio, Texas.
19. *ChemXSeer*. URL: http://cyber-chem.ist.psu.edu/ [accessed: 7/6/2007]
20. Tansley, R., M. Smith, and J.H. Walker, *The DSpace open source digital asset management system: Challenges and opportunities.* Research and Advanced Technology for Digital Libraries, 2005. **3652**: p. 242-253.
21. *DSpace*. URL: http://www.dspace.org/ [accessed: 7/6/2007]
22. Markey, K., et al., *Census of Institutional Repositories in the United States: MIRACLE Project Research Findings*, in *Council on Library and Information Resources*. 2007.
23. *DSpace Federation*. URL: http://wiki.dspace.org/index.php/DspaceGovernance [accessed: 7/6/2007]
24. *Manakin project*. URL: http://di.tamu.edu/projects/xmlui/manakin/ [accessed: 7/6/2007]
25. *Configurable Submission System*. URL: http://hdl.handle.net/2142/207 [accessed: 7/6/2007]
26. *DRIADE Workshop May 2007*. URL: http://driade.nescent.org/Public:DRIADE_Workshop_May_2007 [accessed: 7/6/2007]
27. *OpenID - open, decentralized, free framework for user-centric digital identity*. URL: http://openid.net/ [accessed: 7/6/2007]

28. Greenberg, J., K. Spurgin, and A. Crystal, *Final Report for the AMeGA (Automatic Metadata Generation Applications) Project*. 2005, University of North Carolina at Chapel Hill and Library of Congress.
29. Greenberg, J., K. Spurgin, and A. Crystal, *Functionalities for Automatic-Metadata Generation Applications: A Survey of Metadata Experts' Opinions.* International Journal of Metadata, Semantics, and Ontologies, 2006. **1**(1): p. 3-20.
30. Greenberg, J. and T. Severiens, *DCMI-Tools: Ontologies for Digital Application Description, ELPUB2007*, in *ELPUB2007. Openness in Digital Publishing: Awareness, Discovery and Access - Proceedings of the 11th International Conference on Electronic Publishing*, L. Chan and B. Martens, Editors. 2007: Vienna, Austria. p. 437-444.
31. Anderson, J. and Pérez-Carballo, *The nature of indexing: how humans and machines analyze messages and texts for retrieval. part I: research, and the nature of human indexing.* Information Processing Management, 2001. **37**(2): p. 231-254.
32. Han, H., et al., *Automatic Document Metadata Extraction using Support Vector Machines*, in *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*. 2003, ACM Press: Houston, Texas. p. 37 - 48.
33. Takasu, A., *Bibliographic attribute extraction from erroneous references based on a statistical model*, in *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*. 2003, ACM Press: Houston, Texas. p. 49 - 60.
34. Hatala, M. and S. Forth, *System for Computer-aided Metadata Creation*, in *The Twelfth International World Wide Web Conference*. 2003: Budapest, Hungary.
35. Patton, M., et al., *Toward a metadata generation framework: A case study at the John Hopkins university*. D-Lib Magazine, 2004. **10**(11).
36. *ITIS, the Integrated Taxonomic Information System*. URL: http://www.itis.gov [accessed: 7/9/2007]
37. *uBio - Universal Biological Indexer and Organizer*. URL: http://www.ubio.org [accessed: 7/9/2007]
38. *Lemur Toolkit for Language Modeling and Information Retrieval*. URL: http://www.lemurproject.org/ [accessed: 7/6/2007]
39. *CLAIR (Computational Linguistics And Information Retrieval)*. URL: http://tangra.si.umich.edu/clair/clairlib [accessed: 7/6/2007]
40. *Curation Manual: Open Source for Digital Curation*. URL: http://www.dcc.ac.uk/resource/curation-manual/chapters/open-source/ [accessed: 7/6/2007]
41. *JHOVE - JSTOR/Harvard Object Validation Environment*. URL: http://hul.harvard.edu/jhove/ [accessed: 7/6/2007]
42. *Xena - Digital Preservation Software*. URL: http://xena.sourceforge.net/index.html [accessed: 7/9/2007]
43. Paynter, G., *Developing Practical Automatic Metadata Assignment and Evaluation Tools for Internet Resources*, in *International Conference on Digital Libraries*. 2005, ACM Press: Denver, CO. p. 291 - 300.
44. Altman, M. and G. King, *A Proposed Standard for the Scholarly Citation of Quantitative Data.* D-Lib Magazine, 2007. **13**(3/4).
45. Hilse, H.-W. and J. Kothe, *Implementing Persistent Identifiers*. 2006.
46. *OpenURL*. URL: http://www.exlibrisgroup.com/sfx_openurl_syntax.htm [accessed: 7/6/2007]
47. Apps, A. and R. MacIntyre, *Why OpenURL?* D-Lib Magazine, 2006. **12**(5).
48. Clark, T., S. Martin, and T. Liefeld, *Globally distributed object identification for biological knowledgebases.* Briefings in Bioinformatics, 2004. **5**(1): p. 59.
49. *CNRI's Handle System*. URL: http://www.handle.net [accessed: 7/6/2007]
50. *Digital Object Identifiers*. URL: http://www.doi.org [accessed: 7/6/2007]
51. *caBIO*. URL: http://cabio.nci.nih.gov/NCICB/infrastructure/cacore_overview/caBIO [accessed: 7/6/2007]
52. Good, B. and M. Wilkinson, *The Life Sciences Semantic Web is full of creeps!* Briefings in bioinformatics, 2006. **7**(3): p. 275-86.
53. *STD-DOI Publication and Citation of Scientific Primary Data*. URL: http://www.std-doi.de/ [accessed: 7/6/2007]
54. *OAI Protocol for Metadata Harvesting specification (OAI-PMH)*. URL: http://www.openarchives.org/OAI/openarchivesprotocol.html [accessed: 7/6/2007]

55. *National Science Digital Library (NSDL)*. URL: http://nsdl.org/ [accessed: 7/6/2007]
56. *BioOne*. URL: http://www.bioone.org/ [accessed: 7/6/2007]
57. *Public Library of Science (PLoS)*. URL: http://www.plos.org/ [accessed: 7/6/2007]
58. *OAICat*. URL: http://www.oclc.org/research/software/oai/cat.htm [accessed: 7/6/2007]
59. *OAI-PMH harvester*. URL: http://www.oclc.org/research/software/oai/harvester2.htm [accessed: 7/6/2007]
60. *Metacat*. URL: http://knb.ecoinformatics.org/software/metacat/ [accessed: 7/6/2007]
61. Jones, M., et al., *Managing scientific metadata*. IEEE Internet Computing, 2001. **5**(5): p. 59-68.
62. Berkley, C., et al., *Metacat: A Schema-Independent XML Database System*, in *13th Intl. Conf. on Scientific and Statistical Database Management*. 2001, IEEE Computer Society: Fairfax, Virginia, USA. p. 171.
63. *Ecological Society of America - Data Registry*. URL: http://data.esa.org [accessed: 7/9/2007]
64. *Long Term Ecological Research Network - Metacat Data Catalog*. URL: http://metacat.lternet.edu/knb [accessed: 7/9/2007]
65. *Consortia of ecological research and field stations*. URL: http://knb.ecoinformatics.org/community.jsp [accessed: 7/6/2007]
66. *SRU/SRW (Search and Retrieve via URL (or Web-service))*. URL: http://www.loc.gov/standards/sru/ [accessed: 7/6/2007]
67. *SOAP (formerly for Simple Object Access Protocol)*. URL: http://www.w3.org/TR/soap/ [accessed: 7/6/2007]
68. *RSS (Really Simple Syndication)*. URL: http://www.rssboard.org/rss-specification [accessed: 7/6/2007]
69. *Atom - a universal publishing standard for personal content and weblogs*. URL: http://atomenabled.org/ [accessed: 7/9/2007]
70. *Yahoo Pipes*. URL: http://pipes.yahoo.com [accessed: 7/6/2007]
71. *Creative Commons*. URL: http://creativecommons.org/ [accessed: 7/6/2007]
72. *Science Commons*. URL: http://sciencecommons.org/ [accessed: 7/6/2007]
73. *LOCKSS (Lots of Copies Keeps Stuff Safe)*. URL: http://www.lockss.org/ [accessed: 7/6/2007]
74. *CLOCKSS (Controlled LOCKSS)*. URL: http://www.clockss.org/clockss/Home [accessed: 7/6/2007]
75. *MetaArchive Project*. URL: http://www.metaarchive.org/ [accessed: 7/6/2007]
76. *NESCent's Education and Outreach group*. URL: http://eog.nescent.org [accessed: 7/9/2007]