Project Summary

ABI Development:
Dryad: scalable and sustainable infrastructure for the publication of data

Dryad is a recently launched repository for data underlying the findings in the scholarly bioscience literature. It is distinguished by the close association of data deposition with the process and business of scholarly publishing, and by using article publication as a model for how researchers can benefit from data sharing infrastructure. In a short time, dozens of journals have adopted Dryad as a mechanism for data archiving, and the repository is now at a point of transition to a sustainable organization that has the capacity to make thousands of new datasets each year openly available for reuse in perpetuity.

The **Intellectual Merit** of the proposed work is in the technical and organizational innovations that allow this transition to happen. This will be accomplished by pursuing two aims: scalability and sustainability. The technical goals to ensure scalability include: automation of metadata curation and preservation tasks; developing more efficient and scalable processes to integrate the manuscript submission processes of journals with the data submission process of Dryad; enhancing the features and usability of the deposition interface; and improving the machine and human interfaces for filtering, searching and accessing repository contents. Coupled with these developmental goals, we propose ongoing studies of the costs and benefits of data archiving and data reuse to stakeholders, and continued evaluation of Dryad's role with respect to the many emergent technologies in the world of publishing and data repositories. Organizational goals to transition to sustainability include implementing a nonprofit governance and revenue model that has been developed over the past three years by diverse stakeholders in the research, publishing, library and funder communities.

The **Broader Impacts** of the proposed work are first and foremost the potential to transform scholarly communication within biology. The credibility and effectiveness of the research enterprise can be credited in part to the traditions of scholarly publishing. Researchers are incentivized to disclose their work to their peers in return for professional credit. But in so doing, they expose their findings to be confirmed or refuted, and other researchers may build upon their results. Oddly, data have rarely been disclosed in this way, except for a few common and easily standardized datatypes. New journal and funder mandates, and changes to scientific culture, are creating a demand for new data management, preservation, and dissemination services. Dryad provides a model for how a disciplinary repository can incentivize researchers to disclose the data that is of the greatest value for scientific reuse, that associated with publications, and realize the manifold benefits of free access to scientific data *in perpetuity*. Dryad will actively promote best practices in data archiving and data reuse to the community of current potential users and to the next generation of researchers, through educational initiatives and partnerships with the broader community. The membership of Dryad will provide a forum for participating journals, societies and publishers to take coordinated, and well-informed, steps toward improved practice and standardized policies regarding data.

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.B.2.

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

PROJECT DESCRIPTION

I. Introduction and objectives

The effectiveness of the science is based on no small part on a sometimes implicit social contract under which scientists publish their findings in such a way that they may be confirmed or refuted, and they receive credit for their work in return (Vision 2010). Due perhaps to the limitations of the printed page, and the late arrival of born-digital research artifacts onto the scene, data have been largely left out of this arrangement (Anonymous, 2009). Scientific publications generally contain tables, figures, and statistics based on data, but they do not communicate the curated dataset itself in a way that would allow an independent investigator to validate and extend the work (King 1995).

For certain datatypes, exist specialized public repositories that authors are required to deposit upon publication. The success of specialized repositories such as Genbank is well known (Strasser 2008); whole new disciplines have grown around them. But the vast majority of data types are orphaned -- they do not have such a repository. They exist in what Heidorn (2008) called the "long tail". While there is a huge volume of DNA sequence data, which makes Genbank both so necessary and useful, there may be cumulatively more information content, in an almost limitless variety of form, lurking in the long tail,

Occasionally, orphan data are submitted with articles as part of an online supplement. More often, they are left for authors to share only upon request. Unfortunately, such requests are often left unfulfilled (Savage and Vickers 2009). For example, Wicherts and colleagues (2006) requested data for 141 recent empirical articles in psychology. Though the corresponding authors had signed a certification of compliance with the American Psychological Association, Ethical Principles, which includes the principle of sharing data for independent verification, only one-quarter of the authors furnished their data after six months of repeated requests. Such practices undermine scientific credibility; in a survey of 1240 geneticists, 28 percent reported that they had been unable at some point in the previous three years to confirm published findings because of data withholding (Campbell et al. 2002). Twelve percent admitted that they had not honored at least one request for published data in the preceding three years themselves; the most commonly given reason for denying requests was the amount of effort required for compliance (80 percent of respondents). This is understandable, data files may be misplaced, corrupted, or the software in which they were produced becomes obsolete. Memories of the coding schemes fade. Still, we can and should do better.

The obvious alternative is for authors to archive their processed data, together with sufficient documentation, to an organization that can take responsibility for enabling its discovery, ensuring its preservation, and promoting its dissemination. The archiving of scientific data in this way can offer a host of benefits (Table 1). Recognizing these benefits, there have been an increasing number of calls in recent years to remove barriers to the availability and reusability of scientific data, particularly that associated with publications, and to develop data archiving infrastructures (e.g., Arzberger *et al.* 2004, National Academy of Sciences 2009, Schofield *et al.* 2009, Smith 2009, Guttmacher *et al.* 2009).

It is important to recognize that there are costs to data archiving, as well. In fact, one reason why it is not yet the norm may be that the brunt of these costs are borne by the researchers who must make it happen, yet the benefits to the original researcher are only occasionally realized (e.g. through increased citation, Piwowar 2007). Considerable time is required to curate data to a point where others can use it. Making data available opens up a researcher to potential competition. In the words of Gleditsch and Strand (2003) "will you be scooped or will you be famous?". For these reasons, data archiving may not be possible to achieve without mandates despite its many benefits. Furthermore, funds must be invested in data repository infrastructure, and there is little agreement about which party or parties should shoulder these costs: universities, funding agencies, publishers, etc. By focusing on data associated with publications, at least some of these concerns and uncertainties are alleviated (Costello 2009, Froese et al. 2004). Published data has been through rigorous quality filtering, the publication itself provides detailed documentation to aid in reuse, and authors' proprietary interest in the data is alleviated by virtue of having produced at least the one publication already.

Table 1. Taxonomy of benefits to the archiving of data (after Beagrie *et al.* 2010).

Direct Verification of published research Preserving accessibility to data Allowing reuse and repurposing of data Discoverability of data	**Indirect (costs avoided)** Redundant data collection Inefficient legacy data curation Burden of sharing-upon-request Opportunity cost of science not done
Near term Protection against personnel turnover Availability for review and validation	**Long term** Secure long-term stewardship Increased impact per publication
Private Increased impact and citations Enhanced reader experience New collaborations New research opportunities Fulfillment of funding mandate	**Public** More efficient use of research dollars Public trust in science Educational opportunities Improved methodologies

Journals are in a particularly advantageous position to enact policies that will promote the archiving of data upon publication. The time at which authors are most motivated and able to provide their data in a reusable form is *during* the publication process. There is particular motivation if archiving is a requirement for publication in the journal (as submission of DNA sequence data is for most biology journals). With this in mind, a number of journals in ecology and evolutionary biology have signed on to a joint data archiving policy (JDAP) that requires, as a condition for publication, that the original data reported in an article be archived in an "appropriate public repository" prior to publication (e.g., Whitlock et al. 2010). The policy has built-in flexibility; journals may offer the option of a one-year embargo on public availability. In special circumstances, longer embargoes may be granted at the discretion of an editor, and sensitive information is exempted altogether. The JDAP came into effect simultaneously in January 2011 for some of the leading journals in evolutionary biology (The American Naturalist, Molecular Ecology, Evolution, Journal of Evolutionary Biology, Molecular Biology and Evolution, Heredity, Evolutionary Applications, Evolution), and many other journals, publishers and societies have recently adopted similar policies independently (*e.g.,* Genetics, Science, PLoS), or inspired by the example of the JDAP (*e.g.,* Ecological Monographs, The Journal of Fish & Wildlife Management, The Paleontological Society). At approximately the same time, the National Science Foundation policy that all new proposals must include a Data Management Plan came into effect. This has greatly raised awareness of the issues of data preservation and the accessibility of data not only among researcher, and mobilized libraries, societies, and publishers to find solutions. Realizing a culture of widespread data archiving will require acknowledging the importance of incentives coupled with mandates (Tenopir 2011).

Researchers have several options for archiving data if they are "orphan", in the sense of not being within the remit of a suitable public repository such as GenBank. They may choose to archive their data on a lab or departmental web site that that is managed locally. However, such sites are notoriously unstable (Anderson 2006), and the approach does little to promote standardization, discoverability or long-term preservation. Institutional repositories (Lynch 2003) hold more promise, but to date few institutions have stepped up to the responsibility of stewarding research data to any appreciable degree, and fewer still have taken on the mission of distributing these data to the scientific community and linking them to publications.

Journals may choose either to host archived data themselves through Supplemental Online Materials (SOM) or to direct authors to a suitable public repository. SOM has both strengths and weaknesses for data (Smit 2011). On the plus side, the burden of deposition is light since the author is already submitting the manuscript to the journal. The publisher may be able to provide relatively

informed and inexpensive editorial attention, can ensure a close linkage to the original publication, and has a proven platform for electronic dissemination. On the minus side, the data can seldom be discovered except by manual examination of individual articles. A pay wall sometimes limits access. Publishers put few resources into maintaining SOM and may even fail to migrate it when the journal changes hands. In an eye-opening study, Anderson and colleagues (2006) reported that of the links to supplemental materials ostensibly hosted on the Web site for the journal Genetics, 40% were unavailable just one year after publication. The EU-funded PARSE Insight study found that more than 2/3 of publishers, even large commercial ones, lack any preservation strategy for their SOM, and less than half report providing a mechanism for authors to provide a persistent link to the SOM they host (Smit 2011).

While the practices of publishers around SOM are improving through industry standardization efforts (Carpenter 2010), a dedicated data repository will generally be able to provide a greater level of service than a publisher that handles data as a sideline. Dryad is an example of a public repository that works closely with journals to make data archiving, and responsible data reuse, standard practices within the research community (Vision 2010). Dryad works to achieve this by helping to realize the benefits of data archiving (*e.g.,* enabling data citations) while minimizing its costs (*e.g.,* through integration of manuscript and data submission). Dryad is not just a software platform, but an organization that works to address the full panoply of sociotechnical barriers (legal, financial, etc.) to data archiving and reuse. Dryad aims to be sustainable, with the capacity to ensure long-term preservation and to meet the evolving needs of its stakeholder community by engaging them directly in governance. The current scope of the repository is data associated with published works in the basic and applied biosciences, ranging from environmental science through medicine. While there are many data repositories (Marcial and Hemminger 2010), there are few that operate in a similar manner to Dryad. The closest comparator is Pangaea, which archives georeferenced earth science data, underlies the World Data Center for Marine Environmental Sciences, and has arrangements in place (*e.g.,* with Elsevier) for archiving data as an integral part of article publication (Klump *et al*. 2006).

The Dryad repository has been incubated at the National Evolutionary Synthesis Center (NESCent) in Durham, NC, in collaboration with a group of journals and societies in evolutionary biology and ecology. Representatives from these organizations comprise the Dryad Consortium, which has provided leadership to the effot by setting both organizational and software development priorities. NESCent is an NSF-funded center that is jointly managed by the three Research I universities in the North Carolina Triangle. Each university also has and continues to play an important role in the repository, with software development and community engagement activities based at Duke University, metadata research and curation undertaken through the School for Library and Information Sciences at UNC Chapel Hill, and the production environment managed by the NCSU Digital Library. These three institutions are collaborators on the present proposal, which is being submitted on behalf of the Dryad Consortium in furtherance of the aims of the larger organization.

Here, we propose to build capacity for scaling up the repository so that it has the ability to handle at least an order of magnitude more data on an annual basis, and from dozens to hundreds of journals. This will require working toward greater efficiencies in the journal integration, deposit and curation, systems. To reach this goal, we will also engage the community of researchers, journals, etc. to raise awareness of how Dryad helps to realize the benefits of data archiving and better understand how to address user needs. In parallel, we will evaluate the consequences of repository and journal policies in order to provide evidence for future decision-making and to evaluate the success of the archive. Finally, we will implement a mature revenue plan that will allow Dryad to transition from a research project to an independent not-for-profit organization, one that is governed by its stakeholder community. The goals of this proposal are aligned with Dryad's overall mission and, if successful, this will have a positive transformative impact on the way biological science is conducted.

II. Results From Prior NSF Support

Two projects are especially relevant to the present proposal:

DBI-0743720 "A Digital Repository for Preservation and Sharing of Data Underlying Published Works in Evolutionary Biology**"** $2,180,179 from 9/2008-8/2012, PI: T. Vision (NESCent/Duke), co-PIs J. Greenberg (UNC), K. Antelman (NCSU), W. Piel (Yale U.) and W Michener (UNM). This project, currently in its third year, has provided much of the support for Dryad to date, and led to the public launch of the repository in January 2009. Major development goals included developing a low-burden process of data deposition during manuscript submission in partnership with journals, developing processes for managing identifiers and metadata that promote interoperability, preservation, reuse, and citation; developing mechanisms for exchange of metadata with existing data archives, evaluation and implementation of distributed replication options, and exploration of governance and sustainability models to ensure the long-term viability of the repository. Progress toward these aims is described below. There have been nine publications to date, and over 30 presentations at a wide variety of workshops and conferences. Over ten undergraduate and graduate students, primarily in information science, and one postdoctoral researcher, have received support and training through the project.

OCI-0830944 "DataONE Observation Network for Earth", $501,780 (NESCent subcontract), 08/01/09 - 07/31/14 PI: W. Michener (U. New Mexico), co-PIs: T. Vision (NESCent), S. Hampton (UC Santa Barbara), B. Cook (Oak Ridge National Laboratory). DataONE aims to create an infrastructure of people, technology, and standards to support the full life cycle of biological, ecological, and environmental data and tools, with particular emphasis on connecting data silos, promoting digital preservation, and training a data-literate research community. A set of working groups assists in designing and implementing different aspects of the DataONE cyberinfrastructure (CI), governance, and sustainability models, with broad representation and expertise. The project includes extensive and active community outreach and training components. The core CI consists of three Coordinating Nodes, which handle metadata and network operations, a growing number of Member Nodes (a diverse group of data repositories, including Dryad), and an Investigator Toolkit for discovery, analysis and data management. DataONE is on track to make the first public release of its CI in Fall 2011. A few of the relevant features of DataONE's CI are discussed below.

Repository infrastructure: Dryad is essentially an elaborate customization of DSpace, itself a sophisticated and generic software package for managing digital collections (Smith *et al*. 2003). It is in use by over 800 institutional and other digital repositories around the globe. DSpace was first developed by Hewlett-Packard and MIT and released in 2002. Together with Fedora Commons, DSpace recently came under the umbrella of the nonprofit DuraSpace Foundation. It is an open source project (under a new BSD license) with a extremely active development community numbering in the hundreds.

Dryad takes advantage of DSpace's fundamental agnosticism about the nature of the bit-streams in its collections. In this way, Dryad differs from many other specialized bioscience repositories, which are based on a database with a fine-grained data model. The files of orphan data used by researchers, and hosted by Dryad, may be in effectively any format. This flexibility is critical for Dryad to provide a solution not just for some types of data, or even many types of data, but effectively *any* type of digital data in a researcher's possession. Dryad can be seen as a digital library in which the metadata about the data items are what enable indexing, discovery, management, and reuse.

The DSpace instance at NCSU is the primary node within a multi-tiered replication network designed to ensure redundancy at multiple scales. A Dryad mirror has been established at The British Library (BL) through the DryadUK project, funded by a grant from the JISC. By the time this work would commence, the NCSU and BL sites will provide a first level of failover service, and also reduce network latency for users in the UK and Europe. Already, the NCSU server has 99.7% uptime.

Dryad is a Tier 4 Member Node within the DataONE network (Michener et al. 2010). This means that Dryad will support the full set of DataONE Application Program Interfaces (APIs) and will

participate in distributed replication, both providing and accepting content from compatible Member Nodes and fully supporting a set of content access control rules designed to ensure bit-level integrity for preservation on a decadal timescale.

Before this project would begin, Dryad also plans to become a subscriber to the CLOCKSS service, a preservation archiving service built upon the LOCKSS infrastructure that is used by hundreds of journals as a solution for ensuring dissemination of their articles and supplementary content in then event that they cease operations (Reich and Rosenthal 2009). This provides a backup of Dryad's pages and content to which DOIs would be redirected in the event that Dryad should fold at some point in the future. All content from the archive would then be made freely available. This allows Dryad to provide a reasonable good-faith guarantee of persistent access to the data in snapshot form, on at least a decadal timescale.

Integration of manuscript and data submission: An early priority of the Dryad Consortium was to make the data deposition process extremely simple, so that the JDAP could be reasonably rolled out to all authors, and not just those that are most conscientious. A simple web interface was developed after evaluation of several alternative metadata management and data submission interfaces. The goal was that it should ordinarily take no less than 15 minutes once the files have been prepared, and feedback from depositors suggests that this has been achieved, although controlled timing studies have not yet been done. One of the primary means by which deposition is simplified is that it is integrated with manuscript submission to the journal. The journal provides bibliographic metadata about the submission to Dryad so that a record can be prepopulated with the authors, title, manuscript number, etc. A custom URL is then conveyed by the journal to the corresponding author that points directly to the submission interface at Dryad for that manuscript, using a derivative of the manuscript number as a unique key. The depositor, once logged in, simply uploads their files and is encouraged (but not required) to provide additional metadata to enable reuse (*e.g.,* a separate ReadMe file, an alternative title or author list, subject keywords, a description, etc.). Having the journal convey the Dryad deposit link to the author greatly promotes deposition even for journals where data archiving is not required by policy. However, data may also be submitted through a non-integrated route.

In Dryad's model, all the data files associated with on publication are grouped into a single "data package". Upon deposit, Dryad reserves Digital Object Identifiers (DOI) which will ultimately be registered through the DataCite organization using the California Digital Library EZID service (Michener *et al.* 2011, Starr 2011). DOIs are minted for the data package as a whole and for each separate data file. The data package DOI is conveyed to the journal for inclusion in the published article. Dryad staff, partly manually and partly using automated processes, examine the data files for integrity, curate the metadata for completeness and quality, and may migrate files to preservation formats (see below). The data are released to the public once the article is published online (although the author may choose other options, see below) and the final article metadata (including CrossRef DOI) can be added to the Dryad record. Thus, at the end of the process, there are cross-links between the published article and the data using identifiers that will continue to resolve even in the event of changes to the web address of the article and data.

Dryad uses a relatively lightweight Metadata Application Profile that builds upon existing bibliographic and information standards (White et al. 2008). It is, by design, largely discipline-neutral, but can be expanded to provide disciplinary-specific coverage through control of subject keywords. Furthermore, the article itself provides essential context for reuse. In the words of Eefke Smit (2011): "What better metadata can one think of than the formal research article? Is it not the official version of record, as officially peer reviewed and published, that will explain background, context, methodology and possibilities for further analysis in the best possible way, and express the intentions of the person who helped collect the data?" Dryad's approach is to capture the rich, high-quality metadata that is already produced as part of publication process, and not to unnecessarily burden depositors by expecting or requiring that they provide additional detailed metadata. Discovery of the data in Dryad can be achieved, to a large degree, using existing bibliographic search tools and services.

The first journals to be integrated were mostly, though not exclusively, those that had adopted the JDAP or an equivalent: The American Naturalist, The Journal of Heredity, Evolution, Evolutionary Applications, Journal of Evolutionary Biology, Molecular Ecology/Molecular Ecology Resources, and the Biological Journal of the Linnean Society. More recently, submission integration has been established with Systematic Biology, for which Dryad modified its integration process to make the data securely and anonymously available to peer reviewers. Through the DryadUK project, a number of journals with broader or more biomedically focused content have been recruited, and submission integration is now complete for the first of these, the new open access medical journal BMJ Open. Thus, there are ten journals currently integrated. In addition, there are 20 journals currently in process: including multiple journals from BioMedCentral, Ecological Monographs (as the first of the Ecological Society of America journals to be integrated), Heredity, Integrative and Comparative Biology, the Journal of Fish & Wildlife Management, two journals from the Paleontological Society (Journal of Paleontology and Paleobiology), multiple journals from Pensoft Publishers, and multiple journals from the Public Library of Science (including PLoS ONE). These journals collectively use a wide variety publishing platforms and have divergent submission processes and policies; at this point, we feel we have established proof of concept for the generality of the submission integration approach. However, the effort involved in integrating a new journal (both for Dryad and for the publisher) is a limiting factor, and one we aim to address in the proposed work.

One goal of the first round of NSF funding was to pilot the idea of "handshaking" deposition with specialized repositories. The motivation is that a variety of community-standard repositories provide tools and services which are custom-tailored to particular datatypes (as Genbank is for DNA sequence data). A journal will generally prefer or require that the depositor submit these data to the other repository even in cases where Dryad hosts the source data file. In such cases, a coordinated deposition process could reduce the burden on the user to independently visit multiple repositories to complete a submission, ensure that Dryad does not lower standards when a community accepts the level of curation provided by a specialized repository, and enable the various items in different repositories (and at the publisher site) to be linked to one another automagically. To date, handshaking has been piloted for the two specialized repositories required by the journals currently integrated, TreeBase and GenBank. Handshaking with TreeBase has been in operation since late 2010, and is currently being implemented for GenBank (expected in Fall 2011).

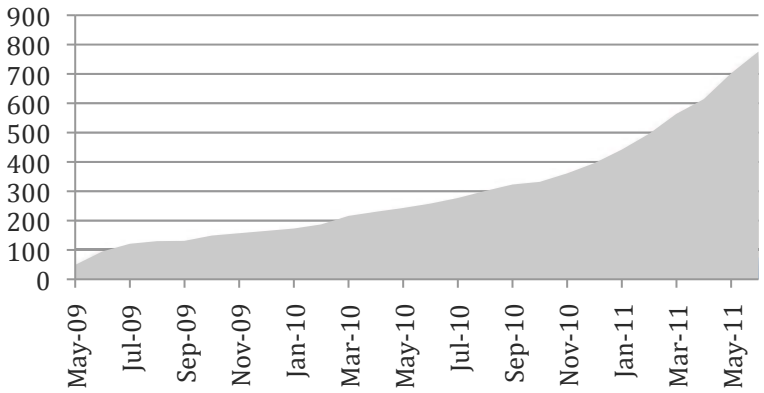
The JDAP allows for data to be embargoed for a period of one year following publication. In order to support this high priority feature, modifications to DSpace were made that would allow authors to select their deposit to be available either at the time of publication (default), immediately upon acceptance by the curator, or following a no-questions-asked one year embargo. Of the thirty journals currently integrated or in process, ten disallow the option of an embargo. Since only three of these journals are currently integrated, the vast majority of authors have been given the option of an embargo upon deposition. Surprisingly, a relatively small proportion of files are given a one-year embargo (9%), despite the no-questions-asked policy. Only 0.5% are under a longer custom embargo, which requires editorial approval; this despite most journals being relatively permissive in granting longer embargoes when requested. Thus, most data files in Dryad are being made available either immediately or, more often, at the time of publication.

This is consistent with our own ongoing research into the risks of being "scooped" by sharing data. Heather Piwowar (unpublished) has examined data reuse citations from one year of deposits to a more established repository, the Gene Expression Omnibus (GEO) and has observed that citations to the data or article that include the original authors decline steeply two years after data deposition, while citations from new researchers begin to accumulate at that time. Given the delay in publication, the period of overlap in reuse by the original investigators and new investigators seems to expire quickly after data deposit, suggesting that the risk of being scooped is relatively small, at least in the realm of gene expression data. Data from an ongoing survey of corresponding authors publishing in JDAP journals also suggests that scooping, while it is perceived to happen with nonzero frequency, is

relatively rare (H. Piwowar, unpublished) despite some researcher concerns to the contrary (Key Perspectives 2010).

A view of the data being deposited: The rate of deposition to the repository is on a healthy upward trend, and has doubled in the past year to its current rate of about 70 per month. As of the end of June 2011, Dryad contained 778 data packages and 1872 data files, associated with 778 articles from 82 different journals (see Figure 1). Since the beginning of 2011, approximately 75% of new deposits have come through the integrated submission process.

Figure 1. *Cumulative submissions to Dryad starting with integration of the first journal (The American Naturalist) in May 2009. The rate of submission continues to grow steadily, currently averaging ~70 data packages per month over the most recent 4-month window.*



On average, data associated with papers in five new journals are deposited each month. The distribution of content by journal is uneven, with the majority coming from integrated journals, but with 35 journals for which there at least two submissions, including many of the most prestigious in biology and science generally. The content is particularly rich in evolution and ecology – not unexpected given the identities of the first integrated journals. For those journals that have been integrated since 2010 and for which the JDAP cam into effect in January 2011, the proportion of articles have data deposited in Dryad from issues in mid-2011 is typically in the range of 40-60%. Some investigators have been going back to old files and systematically uploading data from papers that had previously been published. One investigator has deposited data associated with over 20 papers going back to 2002, most involving many coauthors at different institutions around the globe; no single institutional or national repository had seemed appropriate, until the advent of a global disciplinary repository such as Dryad.

Interestingly, most of the deposits are relatively small in size. Counting all files in a data package together, 80% of the data packages are less than 1 MB in size and 90% are less than 10 MB. A majority of data packages (55%) contain only one data file and 95% contain no more than six data files. Thus, the storage capacity required by Dryad is relatively modest. The repository currently sets an upper limit of 1 GB for each data file and 10 GB for the data package as a whole, but this size limit is rarely approached, and the limit is set more due to issues with network reliability than concerns about storage capacity. It should be noted that since most of these data are processed and hand-curated, they are often quite high in information content even though small in size. On average, authors are providing 2-3 keywords above those provided by the journal (7.6 relative to 5.3), suggesting that many are willing to do minor curation of some metadata given the option, and provided the tools are simple to use.

So far, about 5% of the data packages submitted since handshaking with TreeBase has been available have taken advantage of the feature. This constitutes a non-trivial percentage of the deposits to TreeBase during this time. A very small number of users are currently reporting records that have been independently submitted to other repositories, suggesting a need to rely more on harvesting, and

perhaps greater promotion of handshaking, in order to effectively link data from one publication across the several repositories in which it may exist.

The most recent release of Dryad includes the ability to version of data files and data packages, so that researchers can make updates and corrections. The prior versions are retained "as is" and the user interface displays the relationships among the files, including any description provided for the reason behind the update. Each version receives a unique DOI, and the logical relationship between the different versions, and the different files within a package, are encoded in the DOI for human and machine usability. This provides a unique service to researchers that is typically lacking from Supplementary Online Materials, and indeed many data repositories; it has the potential to lead to a more accurate and less fragmented scientific record. It is too early to know exactly how often this feature will be used, but already curators are handling frequent *ad hoc* requests from depositors for updates to previously deposited data files.

Promoting reuse and data citation: Dryad's terms of reuse are explicit and open in order to reduce the problems that can arise through both ambiguity, and of restrictive proprietary claims of questionable legality that can interfere with legitimate scholarly and scientific use. Authors agree to place their data, and that of their coauthors, into the public domain through use of a Creative Commons Zero waiver. This is the strategy recommended for dissemination of data by Science Commons. In many cases, it is simply stating what is in effect already the case, since facts cannot be copyrighted in many jurisdictions, including the United States. Those who reuse the data are asked to cite the original publication as well as the data package in Dryad, according to scholarly convention, and are provided with guidance and tools to encourage this citation practice. The Dryad Consortium is of the opinion that researchers are incentivized to deposit by the prospect of having their work cited, and that attempting to achieve that by imposing legal restrictions would not be helpful.

The policy of the Dryad Consortium is that both the original article and the data are to be cited when the data are reused as part of a publication. Ensuring that researchers cite the data package in Dryad, and that this citation is tracked, is more complicated than it at first appears (*e.g.,* Parsons and Duerr 2010). It comes as a surprise to many that having a citable DOI does not, in itself, ensure that researchers will receive credit for instances of data reuse. This is partly due to the wholly unstandardized practice on how and where data are cited. Indexing services (*e.g.*, Scopus, Web of Knowledge) do not currently track references to data DOIs either in the text of publications or even, in some cases, when the data citation is included in the references. New standards and services for data citations are being set by DataCite and other organizations, and the infrastructure for data citations is likely to change dramatically in coming years. For now, Dryad does what it can to make data citation as easy as possible by providing its metadata to indexing organizations (the majority of data packages are indexed on Google Scholar), prominently displaying the correct data citation in the user interface, and providing tools for the reference information to be easily downloaded to local reference management software (e.g. EndNote) or uploaded to a social bookmarking and reference service (such as Mendeley).

Usage, so far: As of June 2011, Dryad is averaging 177 unique visitors (and 1,244 pageviews) per day, including visits from 92 countries and every state in the U.S. Presumably many of these individuals are following links directly to data Dryad from the original articles. However, users may also take advantage of Dryad's faceted and filtered search capabilities, and the importance of this interface to users is likely to grow over time. While it would be premature to track published data citations as a measure of the extent of data reuse, it is possible already to examine the number of times data files have been downloaded. Across the repository, the median number of downloads is 10. This is likely an underestimate, as it does not include recently released data files or those still under embargo. Some have been used hundreds of times. The most popular data file has been downloaded over 1000 times (Chave *et al.* 2009, Zanne *et al*. 2009), and one of the first medical datasets from BMJ Open has been

downloaded over 200 times despite having been available for little over a month (Heneghan et al. 2011a,b).

DryadLab: One of the goals of the current NSF funding is to develop and disseminate curricula that use data in the repository that would familiarize future scientists with the reuse of data from Dryad, and in so doing also demonstrate good practices in data management. These are to be distributed under the name of DryadLab. A standard template for DryadLab activities, and a process for design, evaluation, and dissemination has been developed in collaboration with the NESCent Education and Outreach team and the Understanding Evolution Teacher Advisory Group (UE-TAC). A workshop is planned for December 2012 to finalize the first six activities, which have been chosen to represent a diverse set of pedagogically useful studies with data in the repository (including the Zanne dataset). The workshop will bring together a set of the researchers involved in the original publications and members of the UE-TAC, who will work in small researcher-educator groups with Dryad staff to craft the activities and documentation according the template that has been developed. The educators will then test these materials in their own classrooms, with evaluation to be undertaken by NESCent's Education and Outreach staff steering iterative refinement. We intend to continue developing DryadLab activities as an adjunct to the goals described below.

III Development Plan

Clearly, the priorities for development of Dryad needs to be building the capacity to deal with a potentially much larger volume, and diversity, of data deposits, from a larger number of journals, and for building an organization that can sustain the repository for many years past the duration of the proposed project funding. As Dryad transitions to maturity over the next four years, an increasing leadership role will be taken on by the governance structure of the independent not-for-profit organization that is described below, and an increased dependence on revenue sources other than project grants.

III.A Technical Goals

III.A.1 Curation

Currently, curation of each new deposit is estimated to require approximately 30 minutes of curator time. As described below, we are setting as a target a deposit rate of 10,000 new data packages per year, which would require a dedicated curation staff of several FTE. It is important to make the curation process efficient, and at the same time employ automation and new technologies to enrich the quality of the metadata over what can be achieved by manual curation alone. DSpace has relatively immature interface for curation, although it is improving in each new subsequent release. A major goal of this project will be to engineer our own work processes and implement them as features within DSpace or customizations to Dryad. This will be an iterative process, in which the curation feature set will be improved by stages with each release, with field-testing at each stage.

Through funding from the Institute of Museum and Library Sciences to co-PI J. Greenberg at UNC Chapel Hill, we have developed a prototype system named HIVE for enriching subject metadata by text mining of article metadata and data files, and for mapping subject metadata to controlled vocabularies (Greenberg *et al.* 2011). Studies on the accuracy of the keyword suggestions, and the vocabulary mappings, are currently underway. There will need to be extensive parameter tuning and testing before metadata suggestions are displayed in the depositor interface, but we anticipate that we will already have had several months of experience with term suggestions as part of the curator interface by the time this project commences.

One important metadata field that is currently too time-consuming to effectively curate is the author list. Currently, clustering and disambiguation of author names relies on error-prone lexical matches. The Open Researcher & Contributor ID Initiative (ORCID) is an industry-wide effort to establish an open, independent registry for resolving name ambiguity and attribution within scholarly research (Thorisson 2009). The infrastructure is now under active development, and within the next

few years, it is reasonable to expect that the bibliographic metadata for published articles, datasets, and other scholarly works will, as standard practice, include an ORCID for each author. Authors, publishers, research institutions, funders, and scholarly information resources such as Dryad will both contribute and use a common core registry of ORCID profiles. This will provide Dryad with the ability to identify independent works by the same author both within the repository and in external bibliographic resources. Here, we simply propose to pilot the use of these identifiers within Dryad's submission and curation system. We can already anticipate some of the challenges. ORCIDs will be conveyed by journals for each paper, authors may wish to add additional authors to a dataset, requiring functionality for depositors to quickly identify existing ORCIDs for coauthors, or register new ones. There will be a many-to-one mapping between ORCIDs and individuals. The accuracy and completeness of the ORCID database will slowly improve as its used, but may not be initially impressive. Incorporating ORCIDS into existing metadata standards, and web interfaces, will require some experimentation.

Since Dryad has been in operation for less than three years, it has not faced the full force of the challenges in digital preservation. While an initial preservation strategy has been drafted, one that recognizes that file types differ in their ease of preservation, and which provides some guidance to users as to how to deposit their data if they wish to see it reusable decades from now, it has yet to be fully implemented. There has not been sufficient opportunity to test the rate of bit-rot, how reliable it is to pull items out of replicated storage (e.g. DataONE), and the frequency of failure for format migrations. During the course of the proposed work, it will be necessary to improve preservation planning, test the robustness of our processes, and implement a more systematic process for knowing what kind of preservation guarantee can be attached to each kind of file we receive.

III.A.2 Submission Integration

Submission integration is currently accomplished by taking advantage of the email correspondence templates in place with all manuscript processing systems. Messages with the manuscript number and other bibliographic details are automatically sent to, and processed by, Dryad. Templates have been developed for commonly used manuscript processing systems (e.g. ManuscriptCentral from ScholarOne) so that new journals can easily implement the necessary configurations. After deposit, a message is sent back from Dryad containing the data DOI, and the way these are processed by each journal and publisher is highly variable. The way in which publishers disseminate the final information about the article when it is appears is also highly variable. There is room for improvement in all of these steps, to reduce the amount of attention that must be paid by editors to the process, to reduce the time to configure a new journal so that it works properly, to ensure that no manuscript is held up for processing because waiting for data deposition to complete, and to reduce the time between when an article appears and when the data are released within Dryad. We propose to both work with the vendors of these systems to incorporate metadata exchange directly into their products, and to improve Dryad's ability to handle metadata exchange using the technologies already in use by publishers. There is also a need to have additional human capacity for communicating with journals during the submission integration set up process; we have found that with some journals, expertise is limited and it is necessary for Dryad staff to take a very active training role in helping the journal complete the process. As new bibliographic metadata standards such as ORCID are introduced, there will also be a need to update the integration process for all journals. The number of journals is already 30, and is likely to be considerably higher over the coming years, which will necessitate an increased level of coordination and standardization.

III.A.3 Deposition

While Dryad seeks to provide a low-burden deposition interface, it is desirable to support richer standards where possible. One way to achieve this is by enabling journals to require, or users to specify, that certain data should conform to a minimal metadata standard, such as those promoted by the MIBBI project (Taylor et al. 2008). Another is to support packages combining data, software and

analysis steps for replication, as done in Galaxy (Goecks *et al.* 2010) and other workflow systems. Development of flexible interfaces that can accommodate these discipline-specific research outputs will be an ongoing design challenge. Dryad's current policy is to accept data files up to 1 GB in size. While this is not currently a major limitation, we frequently receive requests to host much larger data files. It will be necessary to investigate and possibly fold in mechanisms by which large files can be reliably uploaded (and potentially even replicated) over the network or in the cloud, such as BioTorrents (Langille & Eisen 2010).

It will be necessary to remain flexible to rapid changes in the publishing landscape. A number of publishers are currently in the process of developing data journals, including one recently launched for genome data by BioMedCentral. Others are in the works for biomedicine and biodiversity. Data papers provide a route for researchers to deposit data that might otherwise never be described in the literature, and so represent a potentially important new sector of publishing for a repository such as Dryad. These data publications would likely introduce new requirements for metadata exchange, for curation, and for handshaking with specialized repositories. Other opportunities, such as publication of data associated with PhD theses and conference proceedings are also to be explored.

III.A.4 Search and Retrieval

As Dryad's content grows and becomes more diverse, it will be more of a challenge to offer a view onto the repository that does not overwhelm users with content that is only of marginal interest to them. Some features to be rolled out in the coming year will help, such as landing pages displaying the data packages associated with each journal. However, this does not fully solve the problem of whether and how to filter the content by discipline, such that that someone interested in medical data can have a view of Dryad appropriate to them, while someone interested in evolution can have a different view. It will be necessary to experiment with different approaches to this problem, and to get a greater level of input from users during the design process.

The discovery interface will be revamped to take advantage of the subject metadata and controlled vocabularies that we anticipate through the HIVE project. We will be able to introduce features such as command-line completion for search terms, mapping of synonymies and closely related terms (i.e., recognizing the close relationship when one record has "MHC locus" while another has "HLA locus"), and browsing of hierarchically structure metadata (*e.g.,* taxonomic and geographic names).

Industry leaders in bibliometric indexing such as Thomson Reuters are beginning to track data citations and develop new services around them. At the same time, information scientists are exploring alternative impact metrics and building new services on top of these, as well (http://altmetrics.org/manifesto/). As new services come online for tracking the reuse and impact of individual data packages and data files, we will develop new interfaces to provide these statistics to users.

III.A.5 Design and Development Methodology

The features of the repository will undergo continual evolution following a process of design, prototyping, testing and refinement. Software development goals will be set, and releases are staged, on a quarterly timetable. Requirements for each new feature will be documented on the Dryad wiki and reviewed at monthly all-hands meetings of project personnel. Development tasks will be broken down into units ranging from about 1-8 hours, prioritized and tracked using Fogbugz software. Bug reports and feature requests from users and the project team will also be tracked in Fogbugz, and prioritized at least quarterly. The quarterly releases will be tested by the project team on a development instance of the repository before being moved to staging and production instances.

Standards have been established for user-oriented documentation. Categorized wiki pages are linked from the feature in the release notes. Pages are updated as part of each release and through periodic "documentathons" involving all project personnel. For each feature, the wiki page provides an overview in user-oriented language, step-by-step instructions, links to technical documentation on Google Code, and a design history. Where appropriate, pages include screenshots and workflow

diagrams. Technical features also give the usage, configuration options and an indication of when and how the feature extends or over-rides what is available in DSpace.

User experience based design (UXD, Garrett 2002) is a set of methodologies that seeks to align of system design to the tasks the user seeks to accomplish. Efficiencies in the speed and accuracy of user interactions may come from the design of displays (e.g., the layout of pages, the number of screens), to engineering of the workflow itself, to the architecture of the underlying software. UXD employs such methodologies as workflow analysis, user descriptions, usage descriptions, site maps, content inventories, page wireframes, usability testing, and quantitative analysis of how users navigate the system. An important principle is to guide the design process with early and iterative field observations on user behavior data rather than to rely on survey or focus group results. Since it is critical for Dryad's adoption to have an efficient deposition process, and because expensive human resources must be invested to perform quality control on each deposit, the deposition and curation and interfaces receive dedicated design attention with each release.

III.B. Research and Evaluation

Research and evaluation will enable Dryad to better tune its services to its users and inform how it works with journals to have the greatest impact. One important area of research is reuse analysis. The value of the archive is related to the extent to which archived data is reused for research and education, and to the scientific or educational impact of those reuses. The most common way to measure research impact is through bibliometric analysis of citations. Unfortunately, citations take years to accumulate, making it difficult for depositors to be rewarded, and for the repository to track patterns, on shorter time-scales. Citations also do not capture educational uses. Thus, alternative measures would be desirable. Since usage of Dryad is solely online, access events can be easily measured. To determine if accesses and downloads are predictive of future citations, we will track both direct citations to data, and data reuse instances among citations to the associated article. These will be compared with multiple metrics of access, including aggregated page views and file downloads, patterns of distributions of access over time, number of unique visitors, and method of dataset record discovery (e.g. accession number search, topic search, click-through from the article). We will investigate the extent to which access events and citations are affected by data embargo. We will also attempt to look for patterns that could indicate the level of educational reuse by examination of the timing, and source, of access events. For instance, a cluster of access events from one institution within a short time period suggests usage in a class.

The return on investment for different levels of data curation by either the researcher/depositor or the repository, and the appropriate balance between these two parties is not well understood. Ideally, all data would be provided in standard, preservation-ready formats using machine-readable syntax and semantic standards, come with rich contextual documentation, and so on. Yet, the researcher/depositor - with deep domain knowledge about the data - has other professional tasks competing for his or her time, and the repository - where there is generally good understanding of metadata and data standards - also has limited resources to devote to each deposit. Which aspects of data curation are most correlated with eventual reuse? Can any of the time-intensive tasks be scaled down with little consequence? Are there patterns that would help inform which data deserve extra attention? We will address these questions by examining the data records and reuse patterns of data in a number of repositories including Dryad. We will examine a number of factors that may affect reuse, such as compliance with minimal information standards, availability of raw versus processed data, data formatting, number of controlled/uncontrolled keywords, number and type of files, quantity of free-text documentation, and additional variables to distinguish among different types of data (such as number of authors, subdisciplinary classification, etc).

There are a host of other research questions that can inform operations and policy for the repository and/or the journals with which Dryad works. Many of these lend themselves well to independent research projects for students and summer interns, and we plan to address these as time and resources allow. They include the following. (1) What is the time burden to researchers for depositing data to

Dryad versus sharing data upon request with peers? (2) How does the proportion of eligible articles with archived data in a journal vary with factors such as differences in journal policy, subdiscipline, encouraging authors to use SOM versus a data repository, and the extent to which data deposit fees are borne by the author? (3) How frequently do users navigate from the online journal article to the data package and vice versa, and what proportion of data package discovery is occurring through DOI redirects, through the Dryad search interface, and other routes? (4) Do data that have been made available for peer review differ in the level of researcher-provided curation and reuse?

III.C Communication and Outreach

An important aspect of sustainability is ensuring general awareness of the value proposition of Dryad to its stakeholders, how researchers can take advantage of the repository, and how organizations can participate in its governance, sustainability, and operations. To keep users informed, the repository will continue to post news to its blog on approximately a monthly basis, announcing new features or other developments of interest to Dryad's users and membership. These blog posts are often picked up and disseminated through social media, including Dryad's own Twitter account. Biannual online newsletters will be produced for a wider distribution network. The repository will use a set of mailing lists to allow for two-way conversations with developers, member organizations, deposit purchasers, and integrated journals on topics of particular interest to each community. The repository will develop informational material, both for distribution at conferences (*e.g.,* handouts), and online (*e.g.,* videos) targeted at prospective depositors and data users. Dryad will aim to have a presence at a variety of basic and applied bioscience conferences, through such means such as sponsorships, distribution of promotional materials in registration packets, exhibition booths, and posters. Of course, the greatest promotion is for journals to direct authors to Dryad as a data archiving solution, and for researchers to see Dryad being used for the articles they read – and the data they want to access! We will continue to make sure that institutions are aware of Dryad when training their researchers in data management planning, and to encourage institutions within the scholarly communications community (publishers and others) to recommend Dryad to their authors as a trusted digital repository.

III.D Governance

Dryad has a diversity of stakeholders, including researchers (both those who deposit, and those who reuse), educators and students, scientific societies, publishers, research funders, librarians, data managers, and other consumers and even the larger public that benefits from the increased accessibility and transparency of scientific information. To date, Dryad has been governed by an unincorporated organization called the Dryad Consortium Board, consisting of representatives from Partner Journals.

This Dryad Consortium Board has had three meetings (May 2009 in Durham NC, December 2010 in London UK, and July 2011 in Vancouver BC), during which initial priorities for repository development for set, basic operating principles and policies were established, and issues of governance and sustainability were deliberated. The Board appointed an Executive Committee of five members to provide more regular oversight of repository goals and policies, and to craft specific governance and sustainability proposals for consideration by the Board. The most recent meeting in Vancouver had over 25 representatives from journals, societies and publishers in attendance.

The organization is now in the formative stage of incorporating as a tax-exempt not-for-profit corporation in the state of North Carolina. This step will allow Dryad to raise funds from a greater diversity of donors, to manage its own assets, to oversee operations through contractual arrangements with third parties, and to ensure that the organization stays true to it mission while having the flexibility to adapt to future circumstances. Importantly, it also provides a way for stakeholders, through membership, to have a say in governance that would not be possible were Dryad entirely the responsibility of a host institution. The governance structure that has been developed by the interim Dryad Consortium Board, with the assistance of independent nonprofit counsel, will provide for oversight and strategic decision-making by a Board of Directors (BoD) elected by a broad

organization-based Membership. Here, we provide an outline of the structure as it is currently understood, recognizing that some details may change in the coming months.

Members of the Dryad Consortium are organizations that support its mission and wish to participate in its governance. Any organization that pays an annual fee is eligible to apply for Membership, including but not limited to journals, societies, publishers, research institutions, and funders. Applications for membership are approved by the BoD to ensure that no competing interests have the ability to dominate its numbers. A broad and enfranchised membership will ensure that the organization will remain accountable to its stakeholders. In addition to electing the BoD, members vote on amendments to the Articles and By Laws, and serve more generally as an advisory body to the BoD (through participation in an annual meeting). Membership discounts will be supported for developing countries to the extent possible, and sponsoring memberships will be offered for organizations above the base membership level.

The twelve-person BoD is intended draw upon diverse expertise from the community of stakeholders and beyond. Directors will serve for staggered three terms, and elect Officers of the Board from among their own number. According to standard nonprofit corporate practice, the BoD is vested with legal and financial responsibility to see that the organization fulfills its mission by overseeing its assets and operations. The interim executive committee is charged with soliciting nominations and recruiting the twelve persons who will compose the initial BoD. Once the initial board is in place, a nomination committee will solicit and vet nominations. Directors need not be delegates or represent member organizations, and they may serve sequential terms. Additionally, the board may appoint ex-officio directors, in order to bring in additional expertise.

The executive staff of the repository will report to the BoD directly and are employees of the independent corporation. At least initially, this staff will oversee the contractual arrangement with host institutions (e.g. Duke University). The non-executive repository staff (e.g. software developers, curators), who will execute this project, will be employed by the host institutions. Project coordination will be achieved by monthly all-hands teleconferences, and continued use of the existing communications infrastructure (wiki, mailing lists, etc.) that have been effective to date. The PI and co-PIs of this award will be responsible for coordinating the efforts of the project staff with the BoD and membership (see below) and presenting progress at an annual all-hands meeting.

III.E Revenue

The goal of Dryad's revenue plan is to support an organization that has the resources to manage the data that is deposited today for decades to come, to allow for growth in holdings over time, and to be protected from the vicissitudes of short-term grant funding and the institutional imperatives of any single host institution (Beagrie *et al.* 2009).

The revenue plan is informed by a number of guiding principles: (1) Depositors should be assured that Dryad will have the resources to ensure continued integrity and accessibility of deposited content. (2) Spreading the operating costs for the repository among many contributors reduces the risks of relying on the goodwill of any single organization, and requires only a small contribution from each organization. (3) The costs should also be distributed fairly, in a manner that is proportionate to usage. (4) There should not be charges for access to data, but charges for deposit are acceptable. (5) Diverse organizations should be empowered to pay for data deposits on behalf of researchers, including societies, publishers, research institutions, funding agencies, and individual researchers themselves. (6) Dryad's revenue and expenses should be transparent, and the former should not greatly exceed the latter. More specifically, the revenue model should cover operating costs only, and funds for innovative research and development should be sought from funding agencies and other sources. (7) Revenue will need to scale with the primary cost driver, which is curation at the time of data deposit.

Input from dozens of stakeholder organizations over the past two years have helped to craft a revenue model based on these principles. This model will need to be formally adopted by the BoD before it is official; as it is described here, it reflects the most recent consensus of the (interim) Dryad Consortium. It includes specific price points that assume a certain economy of scale. Currently,

research and development costs are dominant. Subtracting these, the operating costs for Dryad are currently on the order of $100K/yr. If one were to simply divide that by the annual number of new deposits, which at the current rate is on the order of 1K, then $100 would need to be recovered for each deposit. Our aim is for the repository to grow within the next few years to receive 10K deposits annually. This represents less then 1% of the annual number of articles published within Dryad's scope, and – amazingly - less than the annual number of articles published by one of Dryad's current Partner Journals, PLoS ONE. Importantly, while 10K reflects a 10-fold increase in deposits, projections suggest that operating costs would increase by somewhat less than 4-fold. Thus, about $40 per data package is required to cover expenses and this is then our best current guess at the initial baseline deposit charge. It may be adjusted up or down by subsequent growth projections and budget estimates.

Whatever the specific price point, there are several different payment plans through which deposits may be covered: Journal-based, Voucher-based, Pay-as-you-go, and Author pays. Under the Journal-based plan, the journal (or group of journals from one society or publisher) pays to cover any deposit that might be received associated with that journal or journals. To calculate the annual fee, deposit costs are prepaid for the anticipated number of articles with data destined for Dryad. The Voucher-based plan is appropriate for any organization (even a research institution or funder) that wishes to pay in advance for any fixed number of deposits. The Pay-as-you-go plan is similar to the Voucher-based plan, but would allow for deposits to be paid after they occur. Member organizations will be entitled to a discounted rate under the first three plans. Finally, the Author-pays plan is for integrated journals that wish to pass the deposit cost on directly to their authors. There would continue to be an individual-deposit for authors to deposit data associated with non-integrated journals, with a surcharge to over the additional costs of curation. In addition, there would be costs associated with large data, and each payment plan would also include transaction costs, which will favor bulk purchases. Where deposit costs are not covered by another organization, the repository will seek to provide waivers for deposits from authors in developing countries. Growth, expenditure, and revenue numbers and projections will be reviewed at least annually by the Board of Directors (see below) and adjusted as necessary as the sustainability plan is tested by the market.

IV. Broader Impacts

The continued adoption of Dryad repository by a wider community of journals and researchers, and its launch as a self-sustaining organization, will have a transformative impact on scholarly communication across biology. It will allow scholars to contribute to science and receive professional reward for their data collection and curation efforts, for societies and publishers to increase the prestige and impact of the work they publish, and for funders to broaden access to research outputs and allow grant funding to go farther than it currently does (Piwowar et al. 2011). It also stands to improve public trust in science. The science community as a whole will benefit from the availability of data for validation, repurposing, recombination, and value-added data integration products that we cannot yet foresee. The public benefits from having science stand on a firmer foundation, one closer to the virtuous social contract that underpins traditional scholarly communication. And unlike many articles, the data will be in the public domain. A sustainable organization will not only be able provide free access to data in perpetuity, but will also have the resources to subsidize the deposit of data from researchers in the developing world - where a resource-poor bioscience research community is tackling immediate and serious environmental, economic and medical challenges. The repository will serve as a training ground for the entire bioscience community to the extent that Dryad's data, including the DryadLab activities, will be used in classrooms, laboratories and research offices around the world.

References

Anonymous (2009) Data's Shameful Neglect. Nature 461(7263), 461.

Arzberger P, Schroeder P, Beaulieu A, Bowker G, Casey K, Laaksonen L, Moorman D, Uhlir P, Wouters P (2004) An International Framework to Promote Access to Data. Science 303:1777-1778.

Beagrie N, Eakin-Richards L, Vision TJ (2010) Business Models and Cost Estimation: Dryad Repository Case Study, iPRES, Vienna, http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/beagrie-37.pdf

Beagrie N, Lavoie, Woollard M (2010) Keeping Research Data Safe, Phase 2, Final Report to JISC, http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

Blue Ribbon Task Force on Sustainable Digital Preservation (2010) Sustainable Economics for A Digital Plant. Final report. http://brtf.sdsc.edu/

Campbell EG, Clarridge BR, Gokhale M, Birenbaum L, Hilgartner S, Holtzman NA, Blumenthal D. 2002. Data withholding in academic genetics: Evidence from a national survey. Journal of the American Medical Association 287: 473–480.

Carpenter T (2010) Outside the core: working towards an industry recommended practice for supplemental journal materials. Serials 23, 155-158.

Costello, M. 2009. Motivating online publication of data. BioScience 59:418-426.

Froese, R., D. Lloris and S. Opitz. 2004. The need to make scientific data publicly available – concerns and possible solutions. p. 268-271 In M.L.D. Palomares, B. Samb, T. Diouf, J.M. Vakily and D. Pauly (eds.) Fish Biodiversity: Local studies as basis for global inferences. ACP-EU Fisheries Research Report.

Garrett JJ (2002) The Elements of User Experience: User-Centered Design for the Web. Peachpit Press.

Gleditsch, N.P. and H. Strand. 2003. Posting your data: will you be scooped or will you be famous? International Study Perspectives 4:89-97.

Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences/. Genome Biology 11, R86.

Greenberg J, Lossee R, Ramón Pérez Agüera J, Scherle R, White H, Willis C (2011) HIVE: Helping interdisciplinary vocabulary engineering. Bulletin of the American Society for Information Science and Technology 37, 23–26.

Heidorn, P.B. 2008. Shedding light on the dark data in the long tail of science. Library Trends 57:280-

Key Perspectives Ltd. 2010. Data Dimensions: disciplinary differences in research data-sharing, reuse and long term viability. DCC Scarp Synthesis Report. ISSN 1759-586X.

King, G. (1995) Replication, Replication. Political Science and Politics 28(3): 443–499.

- ck, M. Lautenschlager, U. Schindler, I. Sens, and J. Wachter (2006), Data publication in the open access initiative, Data Science Journal 5, 79–83.

Langille MGI, Eisen JA (2010) BioTorrents: A File Sharing Service for Scientific Data. PLoS ONE 5(4): e10071. doi:10.1371/journal.pone.0010071

Lynch, C.A. 2003. Institutional repositories: essential infrastructure for scholarship in the digital age. Libraries and the Academy 3:327-336.

Marcial LH, Hemminger BM (2010) Scientific data repositories on the Web: An initial survey. JASIST 61(10): 2029-2048.

Michener W, Vision T, Cruse P, et al. (2010) DataONE: Data Observation Network for Earth — Preserving Data and Enabling Innovation in the Biological and Environmental Sciences. D-Lib Magazine 17 (1/2), doi:10.1045/january2011-michener

National Academy of Sciences (2009) Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. National Academies Press.

Parsons MA, Duerr R (2010) Data citation and peer review. EOS 91, 297-298.

Piwowar HA, Vision TJ, Whitlock MC (2011) Data archiving is a good investment. Nature 473 (7347), 285.

Piwowar, H.A., R.S. Day and D.B. Fridsma. 2007. Sharing detailed research data is associated with increased citation rate. PLoS ONE 3:e308.

Reich V, Rosenthal D (2009) Distributed Digital Preservation: Private LOCKSS Networks as Business, Social, and Technical Frameworks. Library Trends 57(3), 461-475.

Reichman, O.J., M.B. Jones, and M.P. Schildauer. 2011. Challenges and opportunities to open data in ecology. Science 331, 703-705.

Savage, C.J. and A.J. Vickers. 2009. Empirical study of data-sharing by authors publishing in PLoS journals. PLoS ONE 4: e7078.

Schofield, P.N., J. Eppig, E. Huala, M. Hrabe de Angelis, M. Harvey, D. Davidson, T. Weaver, S. Brown, D. Smedley, N. Rosenthal, K. Schughart, V. Aidinis, G. Tocchini-Valentini and J.M. Hancock. 2010. Sustaining the data and bioresource commons. Science 330:592-593.

Smit E (2011) Abelard and Héloise: Why Data and Publications Belong Together. D-Lib Magazine 17(1/2), doi:10.1045/january2011-smit

Smith, M., Bass, M., McClellan, G., Tansley, R., Barton, M., Branschofsky, M., Stuve, D., Walker, J.H. (2003) DSpace: An Open Source Dynamic Digital Repository. D-Lib Magazine 9(1). doi:10.1045/january2003-smith

Smith, V.S. (2009) Data publication: towards a database of everything. BMC Research Notes 2:113.

Starr J (2011) isCitedBy: A Metadata Scheme for DataCite. D-Lib 17(1/2). doi:10.1045/january2011-starr

Strasser, B. J. 2008. GenBank - Natural history in the 21st century. Science, 322: 537-538. Genbank success - why

Taylor, C.F., D. Field, S.A. Sansone, J. Aerts, R. Apweiler, M. Ashburner, C.A. Ball, P.A. Binz, M. Bogue, T. Booth, A. Brazma, R.R. Brinkman, A.M. Clark, E.W. Deutsch, O. Fiehn, J. Fostel, P. Ghazal, F. Gibson, T. Gray, F. Frimes, J.M. Hancock, N.W. Hardy, H. Hermjakob, R.K. Julian Jr., M. Kane, C. Kettner, C. Kinsinger, E. Kolker, M. Kuip , J. Leebens-Mack, S.E. Lewis, P. Lord, A.M. Mallon, N. Marthandan, H. Masuya, R. McNally, A. Mehrle, N. Morrison, S. Orchard, J. Quackenbush, J.M. Reecy, D.G. Robertson, P. Rocca-Serra, H. Rodriguez, H. Rosenfelder, J. Santoyo-Lopez, R.H. Scheuermann, D. Schober, B. Smith, J. Snape, C.J. Stoeckert Jr., K. Tipton, P. Sterk, A. Untergasser, J. Vandesompele and S. Wiemann. 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nature Biotechnology 26:889-896.

Thorisson, G.A. 2009. Accreditation and attribution in data-sharing. Nature Biotechnology 27:984-985.

Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, et al. 2011 Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6): e21101.

Vision TJ (2010) Open Data and the Social Contract of Scientific Publishing. BioScience 60, 330-330.

White, H.C., S. Carrier, A. Thompson, J. Greenberg and R. Scherle. 2008. The Dryad data repository: a singapore framework metadata architecture in a DSpace environment. Proc. Intl. conf. on Dublin Core and Metadata Applications 157-162.

Whitlock, M.C. 2011. Data archiving in ecology and evolution: best practices. TREE 26: 61-65.

Whitlock, M.C., M.A. McPeek, M.D. Rausher, L. Rieseberg and A.J. Moore. 2010. Data archiving. The American Naturalist 175:145-146.

Wicherts, J.M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. American Psychologist, 61, 726-728.