BUSINESS MODELS AND COST ESTIMATION:

DRYAD REPOSITORY CASE STUDY

Neil Beagrie
Charles Beagrie Ltd
Salisbury, United Kingdom
www.beagrie.com

Lorraine Eakin-Richards
School of Information & Library Science, University of North Carolina
Chapel Hill, NC, USA
www.lorraineeakin.com

Todd Vision
National Evolutionary Synthesis Center
Durham, NC, USA
www.nescent.org

ABSTRACT

Data attrition compromises the ability of scientists to validate and reuse the data that underlie scientific articles. For this reason, many have called to archive data supporting published articles. However, few successful models for the sustainability of disciplinary data archives exist and many of these rely heavily on ephemeral funding sources.

The Dryad project is a consortium of bioscience journals that seeks to establish a data repository to which authors can submit, upon publication, integral data that does not otherwise have a dedicated public archive. This archive is intended to be sustained, in part, through the existing economy of scholarly publishing. In 2009, Dryad commissioned the development of a cost model and sustainability plan. Here we report the outcome of this work to date.

The sustainability efforts of Dryad are expected to provide a model that may be exported to other disciplines, informing the scale needed for a sustainable "small science" data repository and showing how to accommodate diverse business practices among scholarly publishers, funding agencies and research institutions.

1. INTRODUCTION

Researchers, scientists, and publishers recognize that the framework of scientific journal publication is being reinvented as a result of the ascendancy of online access [8]. Sayeed Choudhury of John Hopkins' Virtual Observatory has even argued that the publication of scientific knowledge requires such radically new infrastructure and modes of presentation that while the joint presentation of journals and scientific data may be considered "a new form of compound publication," some situations exist in which "data releases, even without accompanying articles, might be considered a new form of publication" [4]. A host of first mover preservation-oriented organizations and projects such as LOCKSS, Portico, and the DICE group's Storage Research Broker and iRODS infrastructure have set the stage for this new wave of compound publication by creating increased incentives for collaborative preservation of journal articles or research data and by developing solid techniques for ensuring trustworthy preservation. Building from their findings, a number of new initiatives now focus directly upon building the technical and human infrastructure that will enable interoperability between journal articles and associated research data [8][12].

Nonetheless, many of these initial attempts to develop sustainable infrastructure and techniques for creating "enhanced publications,"[1] have focused upon linking large pre-existing databases of research data to the journal articles with which they are associated, such as the Public Library of Science (PLoS) and the Protein Data Bank (PDB) [8][12]. By contrast, the Dryad project focuses on the long tail of datasets reported in the scientific literature that are too heterogeneous in structure to be managed within the (necessarily finite) number of primary bioscience databases.

Because of the frequently more ephemeral and distributed budgetary situations faced by small team scientific publication efforts, the sustainability concerns have required Dryad to develop strategies suited specifically to such efforts. This includes an especially concentrated focus on engaging journal societies and publishers very early in the repository development to ensure that buy-in occurs and is maintained throughout the development cycle. The strategy also includes a large degree of collaboration with institutional partners and like-minded research projects that allow Dryad to take advantage of the inherent cost savings offered by sharing highly skilled personnel and resources.

[1] Woutersen-Windhouwer, S. And Brandsma, R. "Report on Enhanced Publications State-of-the-Art", *DRIVER, Digital Repository Infrastructure Vision for European Research II*, European Union, 2009, p. 7. An enhanced publication is here defined as "a publication that is enhanced with three categories of information: (1) research data (evidence of the research), (2) extra materials (to illustrate or clarify), or (3) post-publication data (commentaries, ranking)." Later in this article, we refer to "supplementary materials and data" to recognize that comparator organizations may include various different scopes of materials when they refer to "data."

© 2010 Austrian Computer Society (OCG).

2. THE DRYAD REPOSITORY

Dryad (www.datadryad.org) is an initiative incubated by The National Evolutionary Synthesis Center (NESCent), the University of North Carolina at Chapel Hill Metadata Research Center, and the North Carolina State Digital Libraries, who began a working consortium of bioscience journals to develop and sustain a digital repository for publication-related data [10]. The repository was initially developed to help support the coordinated adoption of a policy by a number of leading ecology and evolution journals in which data archiving would be required of all authors at the time of publication [11]. Deposition of data into Dryad is one way of satisfying this policy, although other mechanisms are also allowed or encouraged depending upon journal policy (e.g., data may be hosted by the publisher, or archived in specialized repositories such as GenBank). Journals are responsible for making authors aware of their data archiving policy at the time an article is submitted and enforcing it at the time of publication. The repository software is based on DSpace, which allows Dryad to leverage a technology platform being used by hundreds of organizations and maintained by a large and active open-source software community.

Dryad's start-up funds have come primarily from a four year US National Science Foundation (NSF) grant awarded in 2008, as well as NESCent, the NSF-funded DataONE initiative, and the US Institute for Museum and Library Services (IMLS). In addition, a new award through the Joint Information Systems Committee (JISC) in the UK funds Oxford University and the British Library as development partners in Dryad.

The NSF grant identified as a key goal the need to establish stakeholder ownership and governance of Dryad, where journals serve as key stakeholders. To meet this goal, Dryad has created the Dryad Consortium Board (DCB), a central governing body that oversees the repository's strategic planning and to whom the repository staff report. One of the major tasks of the DCB is to agree to a sustainability plan and to help implement it. This will ensure that Dryad can honour its long-term commitment to data preservation.

The DCB currently operates under an interim governance structure consisting of one voting representative from each partner journal. The requirements for partnership for the period prior to the launch of the service in January 2012 [6] include the following:

- Formal adoption of the Joint Data Archiving Policy[2], or an equivalent policy requiring submission of data as a condition of publication;
- Commitment to the development of a self-sustainable business model for Dryad; and
- Appointment of a representative to the DCB with full voting authority.

The DCB elects an Executive Committee of five journal representatives who, together with the Project Director, are responsible for routine oversight of the repository. The Executive Committee is required to bring major financial and governance decisions to the full board for consideration [6].

In addition, partner journals are expected to share article metadata with Dryad prior to publication, to provide information to authors on how to submit to Dryad at the time of submission or acceptance, and to include links to Dryad data within the respective published article, as in [9].

3. A COST MODEL FOR DRYAD

Lorraine Eakin-Richards was commissioned in October 2009 to prepare an initial cost model to help estimate expected repository costs in preparation for sustainability discussions at the DCB meeting in December 2009.

The aim of the cost model was to provide the board with a better understanding of the cost components that the Dryad repository could expect to encounter in both its initial stages of operation and during the early stages of growth expected over a five year time frame. It also provided initial estimates of total and per paper costs.

Eakin-Richards worked with the Dryad project team to assess these current and projected costs, to identify potential cost-share elements from likely ongoing line item budget categories, and to provide a worksheet that could be used on an ongoing basis by the project team to fine tune initial estimates. The key cost components and philosophy of the model were derived, after review of numerous previous cost modeling studies, from the JISC *Keeping Research Data Safe* model [1], which appeared most closely to map to the requirements of the Dryad repository cost modelling needs. The structure of the worksheets was based upon the activity-based cost model built by Eakin and Pomerantz [7].

High level cost categories and potential detailed line item breakdown of these categories were made available to help Dryad begin to project likely costs over time and to fine tune initial estimates as the DCB finalizes its strategies and policies. This breakdown can be viewed in Table 1. Some categories were deemed unnecessary for Dryad's particular situation, such as research and development costs for service innovation, which are expected to be covered via grant funding, and infrastructure costs, which are part of Dryad's institutional partner cost sharing arrangements. Any repository wishing to emulate these categories should

[2]The Joint Data Archiving Policy (JDAP) is a policy of required deposition to be adopted in a coordinated fashion by Dryad partner journals [11]. By adopting the JDAP, a journal agrees to require that data used in support of the conclusions of an article be submitted to a suitable public repository as a condition of publication. Some exceptions hold. For example, authors may elect to embargo access to the data for a period of up to one year following publication of an article; exceptions can also be granted at the discretion of the journal editor in situations such as those in which the data may contain sensitive information regarding human subject data or the location of endangered species.

select those line items most relevant for its own environment and purposes.

One recommendation of the cost modelling consultancy, however, was that a full cost assessment be made and that a risk assessment and strategy be developed to cover the possibility that any particular cost share element be reduced or lost due to budgetary emergencies or strategic changes among the partner institutions. In addition, as longer term planning around operational costs occurs, Dryad will benefit from engaging in time discounting of expenses, in order to gain a better understanding of its "true" long-term economic costs [7].

Repository Management
Repository Manager Salary and Benefits
Advisory Board Meeting Costs
Administrative Support
Administrative Support Salary and Benefits
Curation
Lead Curator Salary and Benefits
Curator Salary and Benefits
Storage and Hardware
System Administrator Salary and Benefits
Hardware Refresh
Security Services
Infrastructure/Facilities
Ongoing Space Expenditures
New Furniture and Equipment Expenditures
Network Set-Up and Maintenance
Telephone and Communications
Research and Development
Personnel Salary and Benefits
Personnel Travel (Specifically Related to Research Collaboration)
Repository Cost Share on Collaborative Projects
Repository Maintenance
Developer Salary and Benefits
Technical Manager Salary & Benefits
Software Expenses
Outreach and Promotion
Communications Specialist Salary and Benefits
Travel for Communications Purposes (e.g., Vendor negotiations, conducting training & workshops)
Advertising Charges
User Documentation and Training
Personnel Salary and Benefits
Outsourcing
Vendor and Consulting Fees
Miscellaneous
Personnel Travel
Personnel Training
Communications Costs (Management, Outreach, Advisory Board, Telephone call charges, etc.)
Miscellaneous Supplies
Insurance
Contingency Estimate

Table 1. Potential Cost Components

The overall projected costs of the service will vary according to the level of investment in value-added services (i.e., data curation). With low to moderate curation effort, initial projections of potential costs for Dryad lead to ballpark estimates of $200,000 or $320,000, respectively, assuming receipt of data from 5,000 or 10,000 papers per annum.

3.1. "Per Paper Costs"

Per paper costs were included within the cost model in order to aid the DCB in determining the feasibility of potential cost recovery techniques. Because buy-in and financial cost sharing from partner journals is a key component of the sustainability model, the journals, societies and publishers needed detailed information about the projected costs, and the repository needed information about what scale of service would be financially viable. The DCB also deemed that a fair model of cost-recovery from journals would need to account for both per paper costs and for the variable number of papers published by each journal in a given year. Given the budget estimates for volumes of 5,000 and 10,000 papers per year, Dryad's per paper expenses were estimated to be $40 and $32, respectively.

3.2. Testing Cost Projections

It is very early in Dryad's development to accurately populate an activity model that could be used to derive full costs for its future activities. In particular, costs for curation will vary according to the level of additional work, e.g., metadata enhancement, and the packaging and documentation for re-use in teaching that may be undertaken by Dryad. Dryad is thus working on the development of a set of "curation service levels" and their associated costs. This is similar to the practice of some publishers, such as the Journal of the American Medical Association or data archives such as the UK Data Archive. Dryad also reviewed use of students or outsourcing to foreign labour markets as part of Dryad's future curation staffing.

4. SUSTAINABILITY PLANNING

In addition to the cost modelling project, Charles Beagrie Ltd was commissioned to work with the Dryad project team to develop sustainability and business planning for the repository. This work began in October 2009 and was completed in April 2010.

The aim was to set a framework in place for future sustainability. Charles Beagrie Ltd incorporated results from Eakin-Richards' cost model and projections by the Dryad project team, led by Todd Vision, within the framework.

The framework is intended to be a dynamic document that can be maintained, reviewed at least annually, and maybe more frequently over the first 2 years, and will evolve over the life of the project and beyond. It provides guidance on sustainability with the aim of

informing business planning. It consists of the following components:

- Strategy, performance indicators and measures
- Comparators and understanding of the costs
- Advantages, benefits and revenue options
- A proposal for sustainability
- Revenue scenarios for Dryad
- Risks register

4.1. External Comparators

We found through desk research and interviews with journals, publishers and data centers that little is known by journals about the specific costs of handling supplementary materials and data. Costs, principally staff time, were observed to vary according to the tasks undertaken and the level of investment in value added services [2]. It is currently not possible directly to compare costs incurred by journals for supplementary data with those for Dryad as they are either largely unknown for the journals, or in Dryad's case, for tasks such as adding metadata to supplementary files, etc., which some journals currently do not undertake. However, it could be observed that the proposed per paper expenses for Dryad appear very reasonable compared to existing author charges (where these exist) for publishing supplementary data files. Amongst three of the journals we interviewed, these author charges for supplementary data files ranged from $100 to $300+.

We also found that while there is no exact archive or repository comparator for Dryad, other archive repositories do offer enough similarities to be of use in comparing some overall costs. Initial analysis, with feedback from the Dryad management team, indicates that a staff of 2-4 FTEs would be a viable initial base level of staffing to deliver Dryad's basic operations. This is comparable to the minimum staffing of other archive comparators at launch we have considered.

The comparators we have reviewed are embedded within larger institutions and can thus leverage pre-existing infrastructure and effort, expertise and direction from associated staff, often co-located but funded separately while working on related activity, including support services, project based research and software development. This helps to maintain a dynamic and sustainable organisation that can respond to change and deal with fluctuations in staffing. Dryad currently is similarly embedded within a larger institution. We noted that Dryad needs to determine the most cost effective way of providing its administration and infrastructure support going forward and ascertain whether support from a host institution can be negotiated at a mutually agreeable cost or provided as an "in-kind" contribution. In due course, a separate not-for-profit legal entity may be considered.

4.2. Transitioning from Project to Service

The transition from Dryad's development phase to a sustainable repository service requires careful planning and the development of a transition strategy. The main considerations revolve around organization and governance; staffing levels; maturity and reliability of automated processes to sustain the repository; and the level of active outreach, training, and member participation to build a critical mass of data available through Dryad. During the transition period, the Dryad team must effectively accommodate changing functional requirements, challenges of scaling the service, and changes in governance. Presently, quarterly repository development plans are reviewed by the Executive Committee, and priorities each quarter are set with careful attention to the needs of current and potential partner journals.

The views of funders on future or continued grant support for Dryad will need to be investigated further. Three categories of grant funding could be important: the possibility of tapered "transition funding" to facilitate the transition from project to service and to allow for the growth of the service in its early years; internationalisation of the service (e.g., mirroring or nodes in Europe or elsewhere) to provide opportunities for widening participation and funding of the service; and research and development opportunities to innovate and enhance the service provided.

Currently 16 interim partner journals participate in the Dryad Consortium. The DCB will consider and agree upon the potential future growth or optimum size of the consortium, appropriate timescales for reaching this size, and impacts on revenues/costs as part of the transition strategy.

5. PROPOSAL FOR SUSTAINABILITY

National or subject repositories are funded in the main through a mixed economy of core and project funding where a maximum of 50% core funding is the norm. Although this can lead to tensions in balancing priorities, diverse revenue streams offer a realistic path for sustaining continued funding and provide some flexibility to decisions around future development.

The easiest and often most successful approaches for projects looking at sustainability issues and possible revenues are to identify those stakeholders that will most benefit from the service and assess their ability and willingness to provide continuing support. Multiple revenue streams can be hard to manage and will bring an additional overhead to the organization that should not be underestimated, so a necessary balance has to be found between the risk of being dependent upon just one or two revenue streams or that of spreading risk across many but then having to deal with managing them.

At the heart of any sustainability model should be a clear articulation of the "value proposition"– how the organization provides a solution to a problem or delivers an attractive product to its stakeholders and users – that would be otherwise difficult, expensive or impossible for them to obtain [3].

For Dryad the value proposition is as follows:

- For scientists, Dryad will increase citations and the impact of their work. It preserves and makes available data that can be used for more complete meta-analysis, for verification of previous results, and to address novel questions with existing data. Dryad provides an easy mechanism for maintaining data over the long term, thereby facilitating compliance with funding agency mandates;
- For publishers, Dryad frees journals from the responsibility and costs of publishing and maintaining supplemental data in perpetuity, and allows publishers to increase the benefits of their journals to the societies and the scientists they support;
- For funding organizations, Dryad provides an extremely cost-effective mechanism for enabling new science and making funded research results openly available.

For future sustainability, the key questions Dryad faces are: what value can be placed on these solutions and products; what are the size and composition of the communities that will receive benefits; and what should be the size of the Dryad Consortium and of the economies of scale delivered to its participants. The feasibility of sustainability for Dryad will depend upon the following factors:

- The costs of maintaining the Dryad organization and its supporting technology;
- The number of partner journals and the rate at which new data packages are ingested;
- Dryad's success in addressing the varying interests of multiple stakeholders, including journals, scientific societies and publishers;
- The extent to which Dryad increases its visibility in the research community, to which there is increase in the practice of data reuse, and to which there is increased adoption of data citations; and,
- The extent to which Dryad can attract funding / revenue for both operating costs and continued development of its service.

Cost and revenue models together with projections and options that should achieve sustainable services over time were presented in a confidential client report. Key components for maintaining the models and sustainability are ongoing review by the Dryad Consortium Board, regular updates to cost and revenue data, and monitoring and updating of the risks register.

6. THE FUTURE

Upon review of the recommendations from the consultancies, the DCB executive committee drafted a prospectus that is currently being circulated among journals, publishers, scientific societies and funding agencies for feedback [5]. The proposal, which will be reviewed by the full DCB in Fall 2010, would establish Dryad as a subscription service by the beginning of 2012. As a major outcome of this work, Dryad is actively expanding the scope of its disciplinary coverage and its institutional partnerships, particularly outside the United States.

7. CONCLUSION

Dryad's funding and development has come at a time when both the sustainability of preservation-oriented programs and the advancement of scientific data repositories has captured the interest of scientists, information science professionals, scientific journals and funding agencies. Dryad's primary aim is to facilitate data discovery and reuse by the research community by ensuring the long-term preservation of the data underlying peer-reviewed articles in the biosciences.

Dryad's business planning efforts are of value beyond the journals and societies directly involved. By testing the idea that both the socio-cultural and economic barriers to data archiving can be overcome within the economy of scholarly communication, Dryad provides a model that may be exported to other disciplines. In particular, it will inform the scope for a sustainable repository, one that balances the need for an economy of scale with the need for cohesion within scientific standards and practices. The model also informs how to accommodate diverse business practices among scholarly publishers, the resources of funding agencies and the capacity of research institutions. To the degree Dryad succeeds in establishing a widely-used and sustainable archive, it will serve as an exemplar for how to realize the full value of the enormous investments in primary scientific data collection.

8. ACKNOWLEGEMENTS

Business planning for Dryad is supported by National Science Foundation grant #0743720. We would like to acknowledge the input of our colleagues Julia Chruszcz, Peggy Schaeffer, and Peter Williams who contributed to the sustainability planning. We would also like to thank the anonymous reviewers of this paper for their most helpful comments and suggestions.

9. REFERENCES

[1] Beagrie, N., Chruszcz, J., & Lavoie, B. *Keeping research data safe: A cost model and guidance for UK universities*, JISC, 2008.

[2] Beagrie, N., Lavoie, B., Woollard, M. *Keeping Research Data Safe 2*, JISC, 2010. Results of the Dryad interviews with publishers will appear in Beagrie, N., Vision, T.J., & Williams, P. *Learned Publishing*, forthcoming.

[3] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access*

to Digital Information, San Diego Supercomputer Center, San Diego, 2010.

[4] Choudhury, G. Sayeed. "The Virtual Observatory Meets the Library," *The Journal of Electronic Publishing*, 11(1). http://dx.doi.org/10.3998/3336451.0011.111.

[5] Dryad Consortium Board Executive Committee. *"Subscription plans"* (proposed). https://www.nescent.org/wg/Dryad/images/8/80/Subscriptions20100422.pdf.

[6] Dryad. *Draft Interim "Governance Plan"*, December, 2009. https://www.nescent.org/wg/Dryad/images/1/1d/Governance_for_dec_meeting.pdf.

[7] Eakin, L. and Pomerantz, J. "Virtual Reference, Real Money: Modeling Costs in Virtual Reference Services", *portal: Libraries and the Academy* 9(1):133-164, 2009. DOI: 10.1353/pla.0.0035.

[8] Fink, J. and Bourne, P. "Reinventing Scholarly Communication for the Electronic Age", *CTWATCH Quarterly* 3(3):26-31. http://www.ctwatch.org/quarterly/articles/2007/08/reinventing-scholarly-communication-for-the-electronic-age/.

[9] Lowry R., Urban, E., and Pissierssens, P., A New Approach to Data Publication in Ocean Sciences, EOS, Transactions American Geophysical Union, 90(50):484-486, 2009. DOI:10.1029/2009EO500004

[10] Vision, T.J. "Open data and social contract of scientific publishing", *Bioscience* 60(5):330-331, 2010. DOI:10.1525/bio.2010.60.5.2.

[11] Whitlock, M.C., McPeek M.A., Rausher M.D., Rieseberg L., and Moore A.J. Data Archiving. American Naturalist 175(2):145-146, 2010. DOI:10.1086/650340.

[12] Woutersen-Windhouwer, S. And Brandsma, R. "Report on Enhanced Publications State-of-the-Art", *DRIVER, Digital Repository Infrastructure Vision for European Research II*, European Union, 2009. http://www.driver-repository.eu/component/option,com_jdownloads/Itemid,83/task,view.download/cid,53/.