Dryad-UK: a repository for data linked to bioscience research articles

"Cherish old knowledge that you may acquire new" The Analects of Confucius

1 Overview

1 The JISC Dryad-UK Project aims to undertake preparative work for the establishment of Dryad as a persistent international repository for bioscience research datasets, linked to journal publication of the peer-reviewed articles they underpin. In particular, it will work directly with the NSF-funded Dryad project that has pioneered this innovative dataset archiving approach, and the international DataCite Association that provides identifiers and services for dataset citation, to archive and publish high-value research datasets that lack natural homes in existing bioinformatics databases.

2 To achieve that aim, this short JISC project has six immediate goals:

2.1 **To create a UK mirror of the Dryad repository**, under the aegis of the British Library (BL), to provide additional dataset security, and yield data for the sustainability business plan.

2.2 **To expand the range of journals** submitting datasets to Dryad, particularly in infectious disease and epidemiology, working with academic publishers, journals and learned societies.

2.3 **To develop a sustainability business plan for the legal, organizational and financial structure of Dryad**, which is currently in a critical phase of transition from prototype to production service, to become a sustainable ongoing international not-for-profit organization empowered to ensure the long-term preservation and accessibility of its data holdings.

2.4 **To promote and facilitate data citation**, by assigning DOIs to Dryad datasets, by developing DaCO, a Data Citation Ontology, and by publishing as Linked Open Data reciprocal citation links between Dryad datasets and the journal articles based upon them.

2.5 **To evaluate the usefulness of Dryad data publication** to the scientific community.

2.6 **To show how Dryad can benefit and support HEIs and funding agencies**, by notifying institutional repositories and funding agencies of datasets published by their researchers.

2 Background

2.1 Context and rationale

1. Imagine a world in which research papers were published on the Web, but in scattered locations, in non-interoperable formats, lacking peer review, without descriptive metadata, and un-indexed by search engines. Worse, imagine that most research papers were not even published in this limited way, but retained within the research groups that created them, only learned about by happenstance, and shared only upon request. Scholarship would be hugely more difficult.
2. For most bioscience research datasets, that *is* the present situation. Once the results from a biological research project have been published in a peer-reviewed journal article, the standard practice is to move on to the next project. The datasets, for which currently the authors can obtain no direct credit, rot quietly in forgotten directories on the hard drives of departed post-docs until all detailed knowledge of their content is lost or forgotten, thus effectively consigning them to oblivion.
3. While many journals require their authors to share data upon request, studies have shown that such requests are frequently denied[1]. And while datasets may be deposited as supplementary information files, these usually lie behind subscription barriers in a variety of soon-to-be-obsolete formats, lack quality control and metadata, have no explicit terms of reuse, and have uncertain permanence[2]. Not indexed by Google, they are discoverable only through their parent articles.
4. Research would be more efficient and cost-effective if those data were preserved in public repositories, made freely available and discoverable, with non-restrictive terms for reuse, persistent identifiers, high-quality metadata and reciprocal links to their parent journal articles. Realization of that vision underlies this proposal.

[1] Campbell *et al*. 2002. Data withholding in academic genetics. *J. Am. Medical Assoc.* 287: 473–480; Wicherts *et al*. 2006. The poor availability of psychological research data for reanalysis. *American Psychologist* 61: 726–728.
[2] Santos et al. 2005. Supplementary data need to be kept in public repositories *Nature* 438: 738; Anderson et al. 2006. On the persistence of supplementary resources in biomedical publications. *BMC Bioinformatics* 7: 260.

2.2 The desirability of the open publication of research data

1. The open publication allows research data to be reviewed and validated, repurposed, included in meta-analyses, and integrated with other data to create new greater wholes. Governments and funding agencies are increasingly voicing their position that the fruits of publicly funded research should be made publicly available in a timely manner, while the academic publisher signatories to the 2007 Brussels Declaration on STM Publishing[3] declared that "... data that are submitted with a paper to a journal should wherever possible be made freely accessible to other scholars".
2. However, in the life sciences, routine archiving of research data is restricted to a few disciplinary islands such as genomics, where well-financed databases exist and where policies of mandatory data submission have been long established. DNA sequence data from EBI have extremely high rates of reuse, demonstrating that public availability of these data is very valuable to the scientific community, as shown by a recent benefits study[4]. And Piwowar[5] found that public availability of gene expression data was associated with a substantial increase in article citations, indicating that researchers also benefit *directly* by sharing their data. Where data sharing is rare, two major reasons are the perceived burden of data submission and the lack of suitable data repositories.[6]
3. The furore surrounding global warming research at the University of East Anglia has focused attention on the need for open data publication. A recent *Nature* editorial[7] points to the importance of the availability of raw data files when evidence in journal articles is challenged. However, authors are frequently unable to find their original data files (Matthew Day, *Nature*, pers. comm.).
4. There are thus many benefits for archiving datasets at the time of publication of journal articles using them. Indeed, an increasing number of academic journals now requiring such data archiving.[8, 9] *Nature*'s policy states that "the preferred way to share large data sets is via public repositories" and provides a list of recommended repositories that includes Dryad.

2.3 What is Dryad?

Dryad (http://datadryad.org/repo) is a DSpace-based digital library for the research datasets that underpin the scientific literature, with links to the articles built upon them, presently funded by the U.S. National Science Foundation (NSF)[10]. Since these data have already demonstrated their value in the hands of the original researchers, there is *a priori* evidence of their potential value to future researchers. Dryad was launched to address the need for a public repository for datasets that had no natural home in more specialized bioinformatics repositories, and was supported by a strong Joint Data Archiving Policy adopted by a group of leading ecology and evolution journals[9]. Its particular strength is allowing authors quickly and easily to deposit datasets of any type (e.g. spreadsheets and multimedia files), through integration with the publishers' article submission processes, enabling preservation and reuse of the 'long tail' of small high-value datasets that underpin the majority of scientific articles. Consistent with the Panton Principles[11], Dryad data are made freely available under the non-restrictive terms of the CC Zero data license[12]. From its initiation, the repository has been governed by an international consortium of partner journals, drawn from a diversity of scientific societies and both commercial and academic publishing houses. Dryad's provision of research data will allow future investigators to validate published findings, explore new analytical methodologies, repurpose the data for research questions unanticipated by the original authors, and perform synthetic studies such as formal meta-analyses.

---

[3] http://www.alpsp.org/ForceDownload.asp?id=304.
[4] Report *Identifying Benefits Arising from the Curation and Open Sharing of Research Data produced by UK Higher Education and Research Institutes*, available at http://ie-repository.jisc.ac.uk/279/, particularly pp 44 ff.
[5] Piwowar *et al*. 2007. Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2: e308.
[6] Smith 2009. Open Access Data publication: towards a database of everything. *BMC Research Notes* 2:113.
[7] Editorial 2010. Under Suspicion. *Nature* 464: 1245.
[8] Editorial 2009. Data's shameful neglect. *Nature* 461: 145.
[9] Moore *et al*. 2010. The need for archiving data in evolutionary biology. *J Evolutionary Biol*. 23: 659-660.
[10] Vision 2010. Open data and social contract of scientific publishing. *Bioscience* 60: 330-331.
[11] http://pantonprinciples.org/.
[12] http://creativecommons.org/publicdomain/zero/1.0/.

3 Appropriateness, Programme Fit, and Value to the JISC Community

3.1 Appropriateness and Fit to Programme Objectives

1. This project is closely aligned with the stated aims of the 14/09 Call of the JISC Managing Research Data Programme, Strand B "Innovative Publications for Research Data". Indeed, Dryad is named as an exemplar of an emerging model of data publication that applications to this call might choose to emulate. We aim to further develop this organizational model for data publication by establishing formal partnerships among the diversity of stakeholders involved in scientific research and scholarly publishing, to create a *self-sustaining* infrastructure. Technical development is targeted specifically to support these organization aims.

3.2 Relationship to partners' current JISC projects

1. The JISC is currently funding the DCC itself, and the following projects led by Dryad-UK partners: LIFE3, NAMES2, ADMIRAL and MILARQ, and the IDMP JISCRDM support project. The proposed Dryad-UK Project has potential synergistic interactions with ADMIRAL, since both are seeking to provide adequate metadata for datasets and both will use BL DataCite[13] DOIs for datasets. Additionally, the PI has made an application to JISC Call 02/10 for the Open Citations Project, which if funded would lead to the publication of article-to-article bibliographic citations as Linked Open Data, giving potential synergies, although no directly overlap, with Dryad dataset-to-article citations. The DCC has evaluated EIDCSR, and similarly has an evaluation role in this project.

3.3 Relationship to other JISC projects and activities

1. The Dryad-UK Project has clear synergies with the WebLinks Project (PI: Brian Matthews, STFC) and parallels with the XYZ Project (PI: Peter Murray-Rust, Cambridge) for which funding applications are being submitted to this same 14/09 call (see reciprocal Letters of Support).
2. Through membership of and active participation in the JISCRDM Programme, we will support, benefit from and build upon the activities of the e-Framework for Education and Research.

3.4 Overall value to the JISC community

1. The JISC has invested heavily in institutional repositories, which Dryad will benefit as follows:
   (a) By trialling an automated notification service to inform the relevant institutional repository of datasets submitted to Dryad by institution members, and by sharing metadata for those datasets for duplicate hosting by the institutional repository, with links to the datasets in Dryad.
   (b) By addressing common standards and ontologies for metadata creation and dataset citation.

3.5 Relationship to other data-centric activities

1. We will continue our close working relationships with the Open Knowledge Foundation, with Science Commons, whose CCZero license is used for Dryad datasets, with DataONE, the NSF-supported cyberinfrastructure for open environmental science data, and with DuraSpace, that provided "open technologies for durable digital content" including DSpace and Fedora, and 'DuraCloud', a cloud infrastructure for hosting open digital content [14]. (See Letters of Support.)
2. We will explore the future integration of Dryad into wider well-funded national and international data infrastructure initiatives that might provide stable long-term support, for example the America Competes Reauthorization Act and the European ELIXIR Framework.

4 Quality of Proposal and Robustness of Workplan

4.1 The Dryad Partnership

1 In forming the Dryad Partnership, we have sought to bring together the major players and stakeholders interested in data preservation. The Dryad-UK Planning Meeting[15], held at HEFCE's London offices on 27-28 April 2010, included practicing epidemiologists, editors and publishers of epidemiological and evolution journals, leaders of the BL, DCC, RIN and UKRDS that have an interest

[13] http://www.datacite.org/.
[14] http://www.okfn.org/, http://sciencecommons.org/, https://dataone.org/, http://duraspace.org/.
[15] https://www.datadryad.org/wiki/Category:DryadUKApril2010.

in the free availability of research data, and representatives of funding agencies that have data preservation activities: the JISC, Wellcome Trust and MRC. (See Letters of Support.)

2 The resulting formal Dryad-UK Partnership, comprises:

2.1 **The British Library**, also representing the DataCite Project and UK Pubmed Central.

2.2 **Oxford University**, representing research biologists whose data we wish to publish, and a leading player in semantic publishing, in JISC RDM, and in the UKRDS Pathfinder Activity.

2.3 The JISC-supported **Digital Curation Centre**.

2.4 **Charles Beagrie**, a consultancy specialising in the digital archive, library and research sectors.

2.5 **The existing Dryad development team** headed by Todd Vision at the National Evolutionary Synthesis Center (NESCent) at the University of North Carolina in the United States.

3. Dryad-UK will be assisted by the following major academic publishers: Biomed Central, Elsevier, Nature, Oxford University Press, PLoS and Wiley-Blackwell. (See Letters of Support)

4. While Dryad-UK will maintain executive autonomy and financial independence from Dryad in the USA during this award, the business and financial plans it produces will be reviewed by the Dryad Consortium Board, with the aim of forming the basis for a sustainable international organization.

4.2 Advantages of the Dryad approach

There are various potential ways to organize data preservation and publication, from voluntary submissions to institutional repositories, supplementary information associated with journal articles. A digital data library such as Dryad can provide numerous advantages over supplementary information attached to journal articles, detailed in [10], including an economy of scale, disciplinary coherence, quality curation, long-term preservation, citability, standardized metadata, improved discovery, and standardized content-delivery mechanisms.

4.3 The importance of the JISC Dryad-UK Project to Dryad

1 Dryad has the potential to transform the landscape of data publication in biosciences, but needs to develop into a financially sustainable organization that is not dependent upon short-term research grants. The work proposed here will address several of the key challenges to enable this:

1.1 The repository needs to develop internationally.

1.2 Dryad needs to expand its disciplinary scope and the number of participating journals, both to achieve economies of scale and to benefit wider sections of the biosciences community.

1.3 Legal, financial and operational details need to be formalized for an international organization that is governed and funded by its stakeholders, is distributed among multiple international institutions, and yet wishes to retain a unitary management structure.

4.4 Workpackages

0. **WP0: Project management** (Lead: BL; involvement: all partners) See **Section 4.5** below.

In addition to management, the work proposed can be divided into six primary workpackages. Most WPs will be conducted in parallel during this short project, thus a GANT chart is unhelpful.

1 **WP1: Establish and use a UK mirror site for Dryad** (Lead: BL, with assistance from NESCent)

We will completely replicate the DSpace backend and Dryad web application. BL and NESCent staff will use this instance to evaluate system requirements to ensure service and data integrity across the distributed system, which in turn will inform the requirements for long-term hosting beyond the term of the project grant. **Effort**: BL 30, NESCent 10 days. **Deliverables**: (i) A pilot service for Dryad-UK, providing a failover front-end and anticipated enhanced response times for UK users through load-balancing (e.g. by origin-of-request). (ii) A report providing technical guidance for a distributed international network with the aim of failsafe data preservation, including investigation of cloud hosting. (iii) Plans and requirements for permanent hosting.

2 **WP2: Expand Dryad journal range and submission rate** (Lead: BL, jointly with NESCent)

Working with publishers and editors who expressed interest during the Dryad-UK planning phase (see Letters of support), and seeking others, we will aim to double the number of journals signed up to Dryad, particularly expanding the area of infectious disease and epidemiology. We will pilot data

submissions from these new journals' authors, and develop relationships with editors and relevant scientific societies to raise awareness. Through the BL's existing contacts, we will raise awareness through presentations, social media channels and other publicity targeted to specific audiences. **Effort**: BL 40, NESCent 10 days. **Deliverables**: New Dryad journals and datasets.

3 **WP3: Formalize partner relations, and develop sustainability plan** (BL, NESCent, Beagrie)

Determining the appropriate legal, financial, managerial, and governance arrangements for the international organization that will oversee the Dryad repository beginning in 2012, after consulting with JISC Legal, institutional representatives and other relevant experts, leading to formalization of these instruments. The work includes preparation of business documents for review by the Dryad Consortium Board; preparation of Memoranda of Understanding among institutional partners; transition planning; constitution of the expanded advisory and governance structure; and implementation of the plans for a mixed funding model (e.g. through subscriptions and submission of a bid to become an ELIXIR node[16]). **Effort**: BL 100, NESCent 10, Beagrie 10 days. **Deliverables:** Sustainability business plan, partner contracts, MOU among host institutions, governance charters, briefing papers for board(s), project management transition plans, potentially one or more applications for long-term sustainability funding.

4 **WP4: Dryad interoperability with publishers and repositories** (Lead: OU, plus BL & NESCent)

4.1 Integration of data with article submission is a critical step in the Dryad process. While the current approach of passing metadata by automated e-mail works well, it requires further engineering to develop templates that will enable participation by a larger number of journals employing manuscript processing systems other than Manuscript Central, and for sending data citation metadata back to journals. **Effort**: OU 30, BL 5, NESCent 5 days.

4.2 A related need is to accommodate 'handshaking' mechanisms to automate data submission and metadata exchange with repositories for special datatypes (e.g. those hosted at EBI), as journals are incorporated that deal with a broader diversity of data types. The data packaging will build upon the work already being undertaken at NESCent for handshaking with TreeBASE and Genbank using SWORD, Bag-It, and ORE[17]. **Effort**: OU 20, BL 10, NESCent 10 days.

4.3 In collaboration with the Oxford University Research Archive (see Rumsey's Letter of Support), we will also prototype harvesting mechanisms through which institutional repositories can be notified of new submissions to Dryad from faculty affiliated with that institution, similar to the service currently offered by BioMedCentral using OAI-PMH[18]. This work will fit well with the proposed UK Research Data Service Pathfinder registry activity at OU. We will also investigate a similar mechanism by which funding agencies can be notified of submissions to Dryad from papers that acknowledge support from that agency, building upon the existing UKPMC Grants Database. **Effort**: OU 40, NESCent 5 days.

Deliverables: (i) Templates for integration of new journals/manuscript submission platforms. (ii) Mechanisms for data deposition to specialist repositories. (iii) Metadata exchange with institutional repositories and funding agencies.

5 **WP5: Metadata standards for data deposition, citation and annotation** (Lead: OU, with BL)

5.1 There is considerable scope for improvement of the metadata standards for datasets and data citations, in order to improve interoperability among publishers and data repositories, and to provide greater syntactical and semantic machine-readability to data citations. OU will lead the development and application of metadata standards for datasets, specifically use of MIIDI[19], the Minimal Information reporting standard for Infectious Disease Investigations for the annotation of new epidemiology and infectious disease datasets. This will include the development of a user-friendly annotation environment using ISA-Creator[20], developed at the EBI and now moving to

[16] http://www.elixir-europe.org/page.php?page=call.
[17] http://www.swordapp.org/, http://www.digitalpreservation.gov/library/resources/tools/docs/bagitspec.pdf, http://www.openarchives.org/ore/, https://www.datadryad.org/wiki/TreeBASE_Submission_Integration.
[18] http://www.biomedcentral.com/info/libraries/oai.
[19] http://imageweb.zoo.ox.ac.uk/wiki/index.php/MIIDI.
[20] http://isatab.sourceforge.net/isacreator.html.

Oxford in conjunction with recently awarded BBSRC ISA Infrastructure grant on which Shotton is Co-I . **Effort**: OU 65 days.

5.2 Following successful development of CiTO, the Citation Typing Ontology[21], which enables encoding of journal citation information as RDF, we will develop DaCO, a parallel Data Citation Ontology, and will use this to permit publication of Dryad data citations as Linked Open Data. This will synergise with the JISC Open Citations Project. We wish particularly to explore reciprocal citations between Dryad and UKPMC. **Effort**: OU 65, BL 5 days.

Deliverables: (i) Use of MIIDI metadata standard and ISA-Creator for annotation of Dryad datasets. (ii) Development of DaCO for data citation; data citations on Web as Open Linked Data.

6 **WP6: Assessment and evaluation** (Lead: DCC, with help from BL and NESCent)

6.1 To achieve buy-in among stakeholders, we need to show that Dryad is more that just another repository, by demonstrating that it enables new research to be conducted. To that end, DCC will establish a framework for evaluation of Dryad data usage. In particular, in light of the Blue Ribbon report[22], we wish to establish the value proposition for saving and publishing article-related datasets (as opposed to primary datasets that may have IPR and ethics constraints to their re-use), the direct benefits that accrue to data depositors, and the relative benefits that accrue from depositing data in different archives (journal supplementary materials, Dryad, institutional repositories or more specialized bioinformatics databases). Additionally, NESCent staff will monitor the rate of data submissions from journals, relative to the number of articles published, and evaluate the effectiveness of the archiving policy at each journal. Dryad has also volunteered to serve as a case study for RIN's proposed study *Research publications and research data*. **Effort**: DCC: 30, BL 20, NESCent 10 days. **Deliverables**: (i) A formative assessment framework for future independent evaluations. (ii) One or more publications on the value of data archives to users and the scientific community. (iii) A planning report allowing Dryad to project the rate of new data submissions over time.

7 **WP7: Dissemination** (Lead: BL, with OU, DCC and NESCent) See **Section 5** below.

4.5 Project Management (WP0)

1. The British Library will recruit a dedicated Project Manager for Dryad-UK, working under the day-to-day supervision of Max Wilkinson and the overall authority of the PI Adam Farquhar. They will maintain the project plan and risk register, managing project resources and local budget, coordinate project partners in execution of the work packages, coordinate meetings of partners and advisory board, submit progress reports to the JISC, and represent the project at JISCRDM meetings. They will maintain a project mailing list and an open Dryad-UK wiki in which he will document the project plans and progress. [**BL effort**: 50 days PM, 10 days other staff]
2. In Oxford, David Shotton will manage his local budget and directly supervise the work of the requested RA on WP4, and also that of Silvio Peroni, an IBRG intern (0.4 FTE additional unfunded effort) who will develop DaCO and other semantic enhancements [**OU effort**: 30 days, PI & RA]
3. The core Dryad-UK teams at BL and OU will be in ongoing asynchronous communication by e-mail among themselves and with the Dryad team in the USA, and will have weekly teleconferences and monthly face-to-face meetings at the British Library or in Oxford. Other partners (DCC, publishers, etc.) will participate in these meetings as appropriate. [**Effort**: NESCent 15, DCC: 3 days]
4. A project advisory board comprising Ian Handel (a research epidemiologist from Edinburgh), Michael Jubb (RIN), Ed Pentz (CrossRef), Theo Bloom (Editor in Chief, PLoS Biology), and Peter Murray-Rust (PI of other JISCMRD projects), will meet twice to provide independent oversight.

Deliverables: Project plan, web site, wiki, documentation and reports.

4.6 Risk analysis

Risk assessment	Probability (1 - 5)	Severity (1 - 5)	Impact (PxS)	Action to prevent / manage risk

[21] http://purl.org/net/cito/, Shotton 2010. CiTO, the Citation Typing Ontology. *J. Biomedical Semantics* (in press).
[22] Blue Ribbon Task Force on Sustainable Digital Preservation and Access, http://www.jisc.ac.uk/brtf.

Problem finding staff	2	5	10	Potential project manager already identified.
Short-term funding causes resignation	3	5	15	Mitigate by thorough documentation of progress and test-led technical development.
Distributed working will not function well	2	4	8	Division of responsibility among researchers, centralized project management, shared management tools (wiki, fogbugz), weekly telecons and multiple face-to-face meetings.
Failure to get take-up by journals and publishers, and to work out a sustainability model for Dryad's future	3	1	3	These issues are principle objective of the project, that will impact the future of Dryad after the project end, but not this project itself. Two major work packages address these issues. Prior research with stakeholders has already informed the framework for the business model.
Authors unwilling to publish data	3	1	3	This too will impact the future of Dryad but not this project. Voluntary Dryad submission rates of 25-50% are already being achieved.
Under-estimation of work load involved	3	3	9	Staged work packages will ensure that the fundamental R&D objectives of the project will be achieved, if not all advanced functionalities.

4.7 Sustainability

A major purpose of this project is to determine an organizational and financial model for the long-term sustainability of Dryad International. One of the principal deliverables of the project will be a financial sustainability plan, anticipated to include variable journal subscriptions as outlined in [23], together with core funding from biological data infrastructure initiatives in the USA and Europe.

4.8 IPR

All the data in Dryad are dedicated to the public domain through the Science Commons CCZero license[12]. Code produced will be released under a BSD open source license either through the DSpace Sourceforge project or the Dryad Google Code project[24]. Scholarly communications arising from this work will be distributed as open access publications.

5 Dissemination and community engagement (WP7)

1. This project has already obtained very wide community engagement, including all the major stakeholders in UK research data management, major academic publishers and editors of some relevant research journals. This will be increased by WP2. Through these journals, we will reach a community of many thousands of potential data providers and data users.
2. Dissemination and publicity for Dryad will be promoted through three workshops towards the end of the project, one hosted by the BL, one by the DCC, and one by NESCent in the USA (at their expense); through papers to international conferences and appropriate journals; and through policy editorials about data archiving in selected journals, e.g. [9]. We will participate as appropriate in relevant community meetings, e.g. DataCite, DCC. Dryad presently maintains a public wiki[25], a blog, a number of mailing lists, and an online presence through a number of other social networking outlets (Facebook and Twitter)[26]. **Effort**: BL 20, DCC 6, OU 10, NESCent 15 days. **Deliverables**: Workshops, journal articles, conference papers and journal editorials, wiki, etc.

6 Impact and beneficiaries

6.1 Benefits to scholarship, teaching and learning, and society

1 In the past, traditional libraries have preserved books in physical form and made them accessible through a card catalogue and open shelves. Analogously, the digital data library of the 21st Century

[23] https://www.nescent.org/wg/dryad/images/8/80/Subscriptions20100422.pdf.
[24] http://sourceforge.net/projects/dspace/; http://code.google.com/p/dryad/.
[25] http://datadryad.org/wiki.
[26] Facebook: http://www.facebook.com/pages/Dryad/; Twitter: http://twitter.com/datadryad, #datadryad.

has the responsibility of preserving high-value research datasets and making them available for discovery and reuse via human and machine-readable online interfaces.

2 This has yet to be achieved for most of research data backing the peer-reviewed scientific literature. Dryad has the potential to address this failing by offering an integrated set of solutions to the technical, socio-cultural, organizational and financial barriers to data publication. While the impact would be dramatic, the chances of success are also high, because the model requires minimal disruption to the culture of scientific research and the business of scholarly publishing, and is potentially extendable to other subject areas and knowledge domains.

3 The beneficiaries of such a repository will be manifold:

3.1 Researchers will be credited with trackable data citations for datasets they have deposited, and will be enabled to do new and better science with the data that are deposited by others.

3.2 Research institutions and publishers will be freed from the burden of long-term data preservation and the provision of an infrastructure for data discovery.

3.3 Research funding agencies can track return on their investments in primary data collection.

3.4 Institutional repositories can be notified of datasets, linked to the journal publication process.

3.5 Open research datasets have enormous potential benefits for teaching and learning.

3.6 Societal benefits relate to public health, biodiversity, and other areas of applied biosciences.

6.2 Benefits to Dryad-UK Partners

1. Having arisen from modest beginnings in a limited subject area, Dryad has the potential to provide significant benefit to the larger international bioscience community. In true Web 2.0 style, it will increase in value as the user-base expands. JISC support of Dryad-UK will facilitate this expansion of Dryad both internationally and into new knowledge domains, to achieve a critical mass of journals and institutional hosts to sustain and preserve the organization, and to increase the usefulness of the repository to the scientific community.
2. This project will enable Dryad to tap the expertise in data management and digital preservation of globally pre-eminent UK organizations, particularly at the BL and DCC. UK involvement is also important because of its centrality to academic publishing, the health of its research community, and the sophistication of its research funding bodies. Involvement in Dryad-UK is central to the mission of the BL and DCC, who will benefit by increasing their leading position in formulating UK data management practices and policies. The University of Oxford will benefit from the synergies between this project and its UKRDS Pathfinder activities, by prototyping links between Dryad and its institutional repository ORA, and by building on its strengths in data semantics.

7 Budget

7.1 Budget justification

1. **Staff**: We request salaries for the **Project Manager and Dryad Developer** at the BL, whose will manage the project and work with other BL staff to draft the business plan, organize mirror hosting, interact with publishers and scientific societies, and engage in evaluation and dissemination; and for the **Interoperability and Semantic Enhancements Developer** at OU, whose work is described in WP4 and WP5. Except for 30 days each for these two staff reserved for personal development and training, e-mails and other time-consuming administrivita, all available staff effort, both funded by JISC and donated to the project, is fully allocated among the WPs detailed above.
2. **Equipment costs**: We request laptop computers for the staff, and half Dryad mirror hosting costs.
3. **Travel expenses**: We request funds to enable regular face-to-face meetings between partners, for attendance at relevant JISC meetings, and for dissemination and sharing of ideas by attendance at two appropriate international conferences (7th Intl. DCC Conf. and ISWC), essential if we are to stay at the cutting edge of data management developments and Semantic Web technologies.
4. **Consumables expenses**: A small 'consumables' budget is requested for computer accessories and printer supplies, and to cover open access journal publication charges.

5. **Dissemination**: We request £3000 towards the costs of the two dissemination / engagement workshops. BL and DCC will each provide a venue for these workshops free to Dryad-UK.
6. **Consultancy**: We request £5000 for 5 days RA and 3.5 days senior staff consultancy effort on the Dryad sustainability model from Charles Beagrie Ltd. Neil B. will donate a further 1.5 days effort.
7. The proposed project leverages a variety of pre-existing US investments in Dryad. NESCent (through funding from NSF and the three universities with which it is affiliated) provides salary release for Todd Vision and supports half the salary of Ryan Scherle, the lead developer on Dryad. The NSF Dryad project grant ($2.18M over four years) provides salary support for technical development, curation, and metadata research, and funds meetings and coordination of the Dryad Consortium Board. The Institute of Museum and Library Sciences ($540K over 3 years) supports the HIVE project[14] (PI: Jane Greenberg, University of North Carolina – see Letter of Support), which provides an interface for metadata enhancement through text mining of Dryad contents and mapping of terms to controlled vocabularies. Dryad is a member of DataONE[14] (a 5 yr, $20M NSF project) which, in addition to technical infrastructure, funds Dr Heather Piwowar to research the socio-cultural incentives and obstacles to data archiving, and the value of publicly available data as a driver to future research. Dryad is also an Associate Partner on the proposed STARSHIP EU Framework 7 project (PI: Matthias Hemmje). Should Dryad be successful in its transition to self-sustainability, the requested JISC funding will leverage substantial *future* sums for repository preservation in the form of subscriptions from partner journals and other sponsors.
8. From its NSF budget, Dryad will cover travel of US participants to Dryad-UK meetings (£4,000), half the cost of Dryad mirror hosting (£4,500), the effort of NESCent personnel (10% effort for R. Scherle, 15% effort for T. Vision, and 20% combined effort for other Dryad staff (content curation, system administration, project coordination, and communications) (£49,000), and £10,000 for an international workshop on metadata standards for data citation. Total contribution value: £67,500.

8 Project Team (abbreviated descriptions due to space limitation)

1. **Dr Adam Farquhar** (Principle Investigator) is Head of Digital Library Technology at BL, was a lead architect on the BL's Digital Library System, co-founded the Digital Preservation Team, and initiated the BL's Dataset Programme. He is Co-ordinator and Scientific Director of the EU co-funded Planets Project and founder of the Open Planets Foundation, and President of DataCite.
2. **Dr Max Wilkinson** (Project Partner) is the Programme Manager for the BL's Dataset Programme, will be line manager for the DRYAD-UK Project Manager, and will facilitate the issue of DataCite DOIs for DRYAD-UK data submissions.
3. **Dr Lee-Ann Coleman** (Project Partner) is Head of Scientific, Technical and Medical Information at the BL. She and her team will seek involvement and engagement with the scientific publishing community and broader stakeholder groups, and will support Dryad-UKPMC interactions.
4. **Kevin Ashley** (Project Partner) is Director of the Digital Curation Centre and the former head of the digital archives department at the University of London Computer Centre, with extensive knowledge of the information management issues facing colleges and universities. He and his DCC staff will evaluate the Dryad-UK Project and assist integration with the JISCRDM Programme.
5. **Dr David Shotton** (PI) heads the Image Bioinformatics Research Group (IBRG) at OU, that publishing and enhancing biological research data by the application of Web and Semantic Web technologies. He has extensive experience in leading JISC projects, interests in data management, , ontologies and semantic publishing, close research links with 'bench' biologists.
6. **Todd Vision** (unfunded International Project Partner) is PI on the NSF grant that currently funds Dryad, Associate Professor in Biology at University of North Carolina at Chapel Hill, and Associate Director for Informatics at NESCent. He has been intimately involved in the planning for Dryad-UK, and will collaborate closely in all aspects of the Dryad-UK project execution.
7. **Ryan Scherle** (unfunded International Project Partner) is the Data Repository Architect at NESCent and lead developer on Dryad since mid-2007. He will integrate code produced by this project into the development and production Dryad instances, assist in the establishment of the UK mirror, and provide advice with regards to all aspects of technical design.

8. **Neil Beagrie** (independent consultant) is head of Charles Beagrie, Ltd., a consultancy company that has already played a key part in sustainability planning for the Dryad repository in the USA, and in the JISC Research Data Management Programme. He will assist Dryad-UK to undertake cost/benefit analyses and establish a sustainability model for future development.